Journal of Machine Learning Research 26 (2025) 1-63 Submitted 12/23; Revised 2/25; Published 9/25

Optimal Rates of Kernel Ridge Regression under Source Condition in Large Dimensions

Haobo Zhang zhang-hb21@mails.tsinghua.edu.cn

Yicheng Li liyc22@mails.tsinghua.edu.cn

Weihao Lu luwh19@mails.tsinghua.edu.cn

Qian Lin* qianlin@tsinghua.edu.cn
Department of Statistics and Data Science
Tsinghua University

Editor: Gabor Lugosi

Abstract

Motivated by studies of neural networks, particularly the neural tangent kernel theory, we investigate the large-dimensional behavior of kernel ridge regression (KRR), where the sample size satisfies $n \asymp d^{\gamma}$ for some $\gamma > 0$. Given a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ associated with an inner product kernel defined on the unit sphere $\mathbb{S}^d$, we assume that the true function f_ρ^* belongs to the interpolation space $[\mathcal{H}]^s$ for some $s > 0$ (source condition). We first establish the exact order (both upper and lower bounds) of the generalization error of KRR for the optimally chosen regularization parameter λ. Furthermore, we show that KRR is minimax optimal when $0 < s \leq 1$, whereas for $s > 1$, KRR fails to achieve minimax optimality, exhibiting the *saturation effect*. Our results illustrate that the convergence rate w.r.t. dimension d varying along γ exhibits a *periodic plateau behavior*, and the convergence rate w.r.t. sample size n exhibits a *multiple descent behavior*. Interestingly, our work unifies several recent studies on kernel regression in the large-dimensional setting, which correspond to $s = 0$ and $s = 1$, respectively.

Keywords: kernel methods, high-dimensional statistics, reproducing kernel Hilbert space, minimax optimality, saturation effect

1. Introduction

The recent studies of neural network theory have sparked a renaissance in kernel methods, as the neural tangent kernel (Jacot et al., 2018) provides a natural surrogate for understanding sufficiently wide neural networks (Arora et al., 2019; Lee et al., 2019; Lai et al., 2023). When the dimension of data is fixed, there has been extensive literature on the generalization behavior of kernel ridge regression (KRR), one of the most widely studied kernel methods (Caponnetto and de Vito, 2007; Fischer and Steinwart, 2020; Cui et al., 2021). Researchers typically characterize the generalization behavior of KRR using two fundamental factors: *capacity condition* and *source condition*. Let $\{\lambda_i\}_{i=1}^{\infty}$ denote the eigenvalues associated with

*. Corresponding author

a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$, the capacity condition assumes that

$$\mathcal{N}_1(\lambda) := \sum_{i=1}^{\infty} \frac{\lambda_i}{\lambda_i + \lambda} \asymp \lambda^{-\frac{1}{\beta}}, \quad \text{as} \quad \lambda \to 0,$$

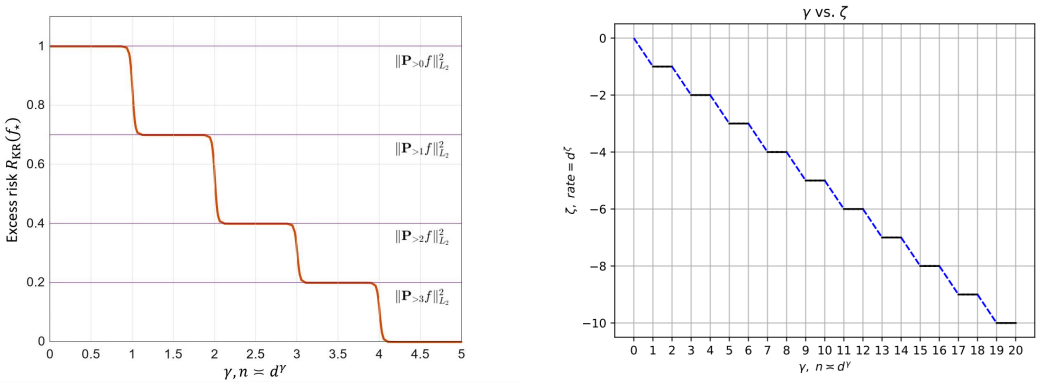
for some $\beta > 1$, where $\lambda > 0$ represents the regularization parameter in KRR. The capacity condition characterizes the size of $\mathcal{H}$ and is frequently stated as an equivalent eigenvalue decay condition: $\lambda_i \asymp i^{-\beta}, \beta > 1$. The source condition assumes that the true function f_ρ^* belongs to the interpolation space $[\mathcal{H}]^s$ for some $s > 0$, i.e.,

$$\left\| f_\rho^* \right\|_{[\mathcal{H}]^s} \leq R,$$

for some constant $R > 0$. It characterizes the relative smoothness of f_ρ^* with respect to $\mathcal{H}$: the larger s is, the "smoother" f_ρ^* is and the easier it can be estimated by KRR. Under this framework, many interesting topics about KRR's generalization behavior were studied. For instance, the minimax optimality of KRR (Fischer and Steinwart, 2020; Zhang et al., 2023) when $0 < s \leq 2$, the saturation effect of KRR (Bauer et al., 2007; Li et al., 2023a) when $s > 2$, the generalization ability of kernel interpolation (Beaglehole et al., 2023; Li et al., 2024), and the learning curve of KRR (Cui et al., 2021; Li et al., 2023b), etc.. We refer to Section 1.1 for further discussion on these topics and detailed explanations of the associated terminologies.

Since neural networks often perform well on data with large dimensionality, studying kernel regression in the large-dimensional setting (where $n \asymp d^\gamma$, $\gamma > 0$) may offer valuable insights into the generalization behavior of neural networks. However, in contrast to the extensive theoretical results in the fixed-dimensional setting, much less is known about the aforementioned topics in the large-dimensional setting. The first obstacle arises from the dependence of RKHS eigenvalues on d in a complex and often unwieldy manner. For example, in the fixed-dimensional setting, the capacity condition actually takes the form: $c(d) \cdot i^{-\beta(d)} \leq \lambda_i \leq C(d) \cdot i^{-\beta(d)}$, where $\beta(d), c(d)$ and $C(d)$ all depend on d, and their explicit expressions can be highly intricate. As a result, the capacity condition does not necessarily hold in the large-dimensional setting (e.g., the inner product kernel on the unit sphere in Section 3.2). Second, we find that $\mathcal{N}_1(\lambda)$ alone is insufficient to establish a tight upper bound of the generalization error convergence rate in large-dimensional setting, marking a key difference from the fixed-dimensional setting. We will see that an extra quantity $\mathcal{N}_2(\lambda) := \sum_{i=1}^{\infty} \left(\lambda_i/(\lambda_i + \lambda)\right)^2$ is needed to characterize the eigenvalues of the RKHS. $\mathcal{N}_1(\lambda)$ and $\mathcal{N}_2(\lambda)$ have the same convergence rate in the fixed-dimensional setting (Remark 2), while the rate of $\mathcal{N}_2(\lambda)$ can be complicated in the large-dimensional setting (Lemma 23).

There are several recent works investigating the generalization error of kernel regression in the large-dimensional setting where $n \asymp d^\gamma, \gamma > 0$. Ghorbani et al. (2021) considers the square-integrable function space on the sphere and proves that when γ is a non-integer, KRR is consistent if and only if the true function is a polynomial with a fixed degree $\leq \gamma$. They also qualitatively reveal that the excess risk exhibits a periodic plateau behavior, cited here as Figure 1(a). Liu et al. (2021) considers the setting $n \asymp d$ and assumes source condition to be $s \in (0, 2]$. They give an upper bound of the generalization error in terms of bias and variance. Using their upper bound, they demonstrate that there could be multiple shapes of the generalization curve as the sample size increases. A more recent work Lu et al.



(a) (b)

Figure 1: Left: Generalization error curve for estimating $f_\rho^* \in L^2$ (borrowed from Ghorbani et al. 2021). Right: The curve of the minimax rates for estimating $f_\rho^* \in \mathcal{H}$ (borrowed from Lu et al. 2023).

(2023) studies early stopping kernel gradient flow in the large-dimensional setting $n \asymp d^\gamma$. Assuming $f_\rho^* \in \mathcal{H}$ and considering the inner product kernel on the unit sphere $\mathbb{S}^d$, they prove an upper bound of the convergence rate and show the minimax optimality of early stopping kernel gradient flow. Interestingly, their results indicate that the minimax rate of the kernel regression for $f_\rho^* \in \mathcal{H}$ exhibits the similar periodic plateau phenomenon, cited here as Figure 1(b). This raises a natural and interesting question: is there a unified way to explain the periodic plateau behavior observed in Ghorbani et al. (2021) and Lu et al. (2023)?

Suppose that $\mathcal{H}$ is an RKHS associated with an inner product kernel defined on the unit sphere $\mathbb{S}^d$. The main focus of this paper is to derive the matching upper and lower bounds of the generalization error and discuss the minimax optimality of KRR for general source condition $s > 0$, i.e., when the true regression function satisfies $f_\rho^* \in [\mathcal{H}]^s$. Allowing s to vary not only represents a more reasonable assumption for the true function, but also provides a natural framework to clarify the relation between the results in Ghorbani et al. (2021) and Lu et al. (2023). In fact, an application of interpolation space theory suggests that the results in Ghorbani et al. (2021) and Lu et al. (2023) are two special cases of our results corresponding to $s = 0$ and $s = 1$ respectively. Generally, this paper has the following contributions:

- We consider a more general framework than the traditional capacity-source condition. We introduce $\mathcal{N}_1(\lambda), \mathcal{N}_2(\lambda), \mathcal{M}_1(\lambda)$ and $\mathcal{M}_2(\lambda)$ in (4), which are key quantities depending on the RKHS, the true function and the regularization parameter λ in KRR. Under mild assumptions, we use these key quantities to express the matching upper and lower bounds of the generalization error as long as the regularization parameter satisfies some approximation conditions (Theorem 1). This framework imposes minimal assumptions on RKHS eigenvalues and the true function, making it applicable to the large-dimensional setting and general source condition. In the fixed-dimensional setting, our results in

Theorem 1 also recovers the state-of-the-art theoretical results about the convergence rates of KRR in Li et al. (2023b).

- We then add source condition into our new framework and consider the inner product kernel on the unit sphere $\mathbb{S}^d$. When $n \asymp d^\gamma$, we derive exact convergence rates (both upper and lower bounds) of the generalization error under the best choice of regularization parameter for any source condition $s > 0$ and almost all $\gamma > 0$ (Theorem 4 for $s \geq 1$ and Theorem 5 for $0 < s < 1$). We will see that the convergence rate w.r.t. dimension d varying along γ exhibits a *periodic plateau behavior*, and the convergence rate w.r.t. sample size n exhibits a *multiple descent behavior*.

- For the inner product kernel on the unit sphere $\mathbb{S}^d$, we further derive the corresponding minimax lower bound for all $s > 0$ and $\gamma > 0$. When $0 < s < 1$, the exact rates in Theorem 5 match the minimax lower bound, and thus we prove the minimax optimality of KRR. When $s > 1$, the KRR is not minimax optimal, i.e., we discover a new version of the saturation effect of KRR (the phenomenon that KRR is not minimax optimal, see the discussion on page 15). In the fixed-dimensional setting, the saturation effect of KRR only happens when $s > 2$. In the large-dimensional setting, we find that a similar phenomenon also happens for $1 < s \leq 2$. Specifically, for any $s > 1$, there are corresponding ranges of γ such that the convergence rates of KRR can not achieve the minimax lower bound even under the best choice of regularization parameter.

1.1 Related work

In the introduction, we highlighted several interesting topics related to the generalization behavior of KRR. These topics have been well-studied in the fixed-dimensional setting. The first fundamental question concerns the minimax optimality of KRR. Under the framework of capacity condition and source condition, Caponnetto and de Vito (2007) proves the minimax optimality of KRR when the source condition satisfies $1 \leq s \leq 2$. Then, extensive literature (see Steinwart et al. 2009; Lin et al. 2018; Fischer and Steinwart 2020; Zhang et al. 2023, 2024 and the reference therein) studies the mis-specified case ($0 < s < 1$), where Zhang et al. (2024) proves the minimax optimality for all $0 < s \leq 2$ under further embedding index condition. The second question concerns the *saturation effect* of KRR, which occurs when $s > 2$. In this regime, no matter how carefully KRR is tuned, the convergence rate can not achieve the minimax lower bound. The saturation effect is conjectured by Bauer et al. (2007); Gerfo et al. (2008) and rigorously proved by Li et al. (2023a). A third area of interest is the generalization ability of *kernel interpolation* (i.e., taking $\lambda = 0$ in KRR), motivated by the remarkable performance of overparameterized neural networks. The results in Rakhlin and Zhai (2019), Buchholz (2022), Beaglehole et al. (2023), and Li et al. (2024) imply that kernel interpolation can not generalize in the fixed-dimensional setting. Last but not least, Bordelon et al. (2020), Cui et al. (2021), and Li et al. (2023b) study the *learning curve* of KRR, which aims to obtain the precise formula (or exact convergence rate) of generalization error for any regularization parameter $\lambda > 0$.

In the large-dimensional setting, these questions remain largely unresolved. Many researchers have studied these problems from different angles and settings. A line of work uses the tools of high-dimensional kernel random matrix approximation from Karoui (2010)

and studies the generalization ability of kernel interpolation (Liang and Rakhlin, 2020; Liang et al., 2020). When $n \asymp d$, Liang and Rakhlin (2020) gives an upper bound of the generalization error of kernel interpolation and claims that the upper bound tends to 0 when the data exhibits a low-dimensional structure. Further, when $n \asymp d^\gamma, \gamma > 0$, Liang et al. (2020) gives an upper bound with a specific convergence rate, which implies that kernel interpolation can generalize if and only if γ is not an integer. One closely related topic is the benign overfitting phenomenon, which we refer to Bartlett et al. (2020), Muthukumar et al. (2020), Hastie et al. (2022), and Tsigler and Bartlett (2023).

Another line of work follows Ghorbani et al. (2021), which has been mentioned in the introduction. This line of work adopts the square-integrable assumption of the true function and aims to obtain the exact generalization error of kernel methods in various settings (Ghorbani et al., 2020; Mei et al., 2022; Mei and Montanari, 2022; Ghosh et al., 2021; Xiao et al., 2022; Hu and Lu, 2022; Misiakiewicz, 2022; Donhauser et al., 2021). To our knowledge, Lu et al. (2023) is the only study that establishes minimax optimality results for specific kernel methods. As discussed in the introduction, Lu et al. (2023) considers the $s = 1$ case ($f_\rho^* \in \mathcal{H}$) and kernel early stopping gradient flow. We will provide a detailed discussion on Ghorbani et al. (2021) and Lu et al. (2023) in Section 4.

2. Preliminaries

Let a compact set $\mathcal{X} \subseteq \mathbb{R}^d$ denote the input space and $\mathcal{Y} \subseteq \mathbb{R}$ denote the output space. Let $\rho = \rho_d$ be an unknown probability distribution on $\mathcal{X} \times \mathcal{Y}$ and denote the corresponding marginal distribution on $\mathcal{X}$ by $\mu = \mu_d$. We use $L^p(\mathcal{X}, \mu)$ (in short L^p) to represent the L^p-spaces. Denote the conditional mean as

$$f_\rho^*(\boldsymbol{x}) = f_{\rho_d}^*(\boldsymbol{x}) \coloneqq \mathbb{E}_{\rho_d}[\, y \mid \boldsymbol{x} \,] = \int_{\mathcal{Y}} y \, \mathrm{d}\rho_d(y|\boldsymbol{x}).$$

Throughout the paper, we make the following assumption:

Assumption 1 *Suppose that $\mathcal{H} = \mathcal{H}_d$ is a separable reproducing kernel Hilbert space (RKHS) on $\mathcal{X} \subset \mathbb{R}^d$ with respect to a continuous kernel function $k = k_d$ satisfying*

$$\sup_{\boldsymbol{x} \in \mathcal{X}} k_d(\boldsymbol{x}, \boldsymbol{x}) \le \kappa^2,$$

where κ is an absolute constant.

For example, in Section 3.2, we consider k_d as the inner product kernel on the unit sphere, which takes the form $k_d(\boldsymbol{x}, \boldsymbol{x}') = \Phi(\langle \boldsymbol{x}, \boldsymbol{x}' \rangle)$, where $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{S}^d$ and $\Phi(\cdot)$ is a function independent of the dimension d (see Assumption 4).

By convention in statistical learning, we consider the large-sample-size limit, i.e., the case where the sample size n is sufficiently large, although our results are non-asymptotic. This paper allows the dimension d to grow as n grows, meaning that we consider $d = d(n)$ and a sequence of estimation problems:

$$\text{Estimating } \{f_{\rho_d}^*\}_{d \ge 1} \text{ using the RKHS } \{\mathcal{H}_d\}_{d \ge 1} \text{ or kernel } \{k_d\}_{d \ge 1}.$$

In Section 3.1, we will consider the general case $d = d(n)$, where d can be fixed or vary with n. In Section 3.2 and Section 3.3, we consider a specific large-dimensional setting, where $n \asymp d^\gamma$ for some $\gamma > 0$. In both cases, we suppose that Assumption 1 holds uniformly for all $d \geq 1$. For brevity, in the remainder of this paper, we frequently omit the subscript d in $\mu_d, \rho_d, f^*_{\rho_d}, k_d, \mathcal{H}_d$, and other quantities that depend on d.

Suppose that the samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ are i.i.d. sampled from ρ. Kernel ridge regression (KRR) constructs an estimator $\hat{f}_\lambda$ by solving the penalized least squares problem

$$\hat{f}_\lambda = \arg\min_{f \in \mathcal{H}} \left(\frac{1}{n} \sum_{i=1}^n \left(y_i - f\left(\boldsymbol{x}_i\right) \right)^2 + \lambda \|f\|_{\mathcal{H}}^2 \right),$$

where $\lambda > 0$ is referred to as the regularization parameter.

Denote the samples as $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$ and $\mathbf{y} = (y_1, \cdots, y_n)^\top$. The representer theorem (see, e.g., Steinwart and Christmann 2008) provides an explicit formula for the KRR estimator:

$$\hat{f}_\lambda(\boldsymbol{x}) = \mathbb{K}(\boldsymbol{x}, \boldsymbol{X})(\mathbb{K}(\boldsymbol{X}, \boldsymbol{X}) + n\lambda \mathbf{I})^{-1} \mathbf{y}, \tag{1}$$

where

$$\mathbb{K}(\boldsymbol{X}, \boldsymbol{X}) = \left(k\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)\right)_{n \times n}, \quad \mathbb{K}(\boldsymbol{x}, \boldsymbol{X}) = \left(k\left(\boldsymbol{x}, \boldsymbol{x}_1\right), \cdots, k\left(\boldsymbol{x}, \boldsymbol{x}_n\right)\right).$$

We aim to analyze the convergence rate of the generalization error (excess risk) of $\hat{f}_\lambda$:

$$\mathbb{E}_{\boldsymbol{x} \sim \mu}\left[\left(\hat{f}_\lambda(\boldsymbol{x}) - f^*_\rho(\boldsymbol{x}) \right)^2 \right] = \left\| \hat{f}_\lambda - f^*_\rho \right\|^2_{L^2(\mathcal{X}, \mu)}.$$

Notations. We use the standard asymptotic notations $O(\cdot)$, $o(\cdot)$, $\Omega(\cdot)$, and $\Theta(\cdot)$. We also write $a_n \asymp b_n$ for $a_n = \Theta(b_n)$; $a_n \lesssim b_n$ for $a_n = O(b_n)$; $a_n \gtrsim b_n$ for $a_n = \Omega(b_n)$; $a_n \ll b_n$ for $a_n = o(b_n)$. We will also use the probability versions of the asymptotic notations such as $O_{\mathbb{P}}(\cdot), o_{\mathbb{P}}(\cdot), \Omega_{\mathbb{P}}(\cdot)$, and $\Theta_{\mathbb{P}}(\cdot)$. For instance, we say the random variables X_n, Y_n satisfying $X_n = O_{\mathbb{P}}(Y_n)$ if and only if for any $\varepsilon > 0$, there exist a constant C_ε and N_ε such that $P\left(|X_n| \geq C_\varepsilon |Y_n|\right) \leq \varepsilon, \forall n > N_\varepsilon$.

2.1 Integral operator and interpolation space

Let $S_k : \mathcal{H} \to L^2(\mathcal{X}, \mu)$ denote the natural embedding operator. Its adjoint operator $S_k^* : L^2(\mathcal{X}, \mu) \to \mathcal{H}$ is an integral operator. Specifically, for $f \in L^2(\mathcal{X}, \mu)$ and $\boldsymbol{x} \in \mathcal{X}$,

$$\left(S_k^* f\right)(\boldsymbol{x}) = \int_{\mathcal{X}} k\left(\boldsymbol{x}, \boldsymbol{x}'\right) f\left(\boldsymbol{x}'\right) \mathrm{d}\mu\left(\boldsymbol{x}'\right).$$

Under Assumption 1, S_k and S_k^* are Hilbert-Schmidt (HS) operators and therefore compact. Their HS norms (denoted as $\|\cdot\|_2$) satisfy

$$\|S_k^*\|_2 = \|S_k\|_2 = \|k\|_{L^2(\mathcal{X}, \mu)} := \left(\int_{\mathcal{X}} k(\boldsymbol{x}, \boldsymbol{x}) \mathrm{d}\mu(\boldsymbol{x}) \right)^{1/2} \leq \kappa.$$

Next, we define two integral operators:

$$L_k := S_k S_k^* : L^2(\mathcal{X}, \mu) \to L^2(\mathcal{X}, \mu), \qquad T := S_k^* S_k : \mathcal{H} \to \mathcal{H}. \tag{2}$$

L_k and T are self-adjoint, positive-definite and trace class (hence Hilbert-Schmidt and compact). Their trace norms (denoted as $\|\cdot\|_1$) satisfy

$$\|L_k\|_1 = \|T\|_1 = \|S_k\|_2^2 = \|S_k^*\|_2^2 .$$

The spectral theorem for self-adjoint compact operators yields that there is an at most countable index set N, a non-increasing summable sequence $\{\lambda_i\}_{i\in N} \subseteq (0,\infty)$ and a family $\{e_i\}_{i\in N} \subseteq \mathcal{H}$, such that $\{e_i\}_{i\in N}$ is an orthonormal basis (ONB) of $\overline{\operatorname{Ran} S_k} \subseteq L^2(\mathcal{X},\mu)$ and $\{\lambda_i^{1/2} e_i\}_{i\in N}$ is an ONB of $\mathcal{H}$. Further, the integral operators can be written as

$$L_k = \sum_{i\in N} \lambda_i \langle \cdot, e_i\rangle_{L^2} e_i \quad \text{and} \quad T = \sum_{i\in N} \lambda_i \left\langle \cdot, \lambda_i^{1/2} e_i \right\rangle_{\mathcal{H}} \lambda_i^{1/2} e_i.$$

We refer to $\{e_i\}_{i\in N}$ and $\{\lambda_i\}_{i\in N}$ as the eigenfunctions and eigenvalues. Mercer's theorem (see, e.g., Steinwart and Christmann 2008, Theorem 4.49) states that

$$k\left(\boldsymbol{x},\boldsymbol{x}'\right) = \sum_{i\in N} \lambda_i e_i(\boldsymbol{x}) e_i\left(\boldsymbol{x}'\right), \quad \boldsymbol{x},\boldsymbol{x}' \in \mathcal{X},$$

where the convergence is absolute and uniform in $\boldsymbol{x},\boldsymbol{x}'$.

Since we are going to consider the source condition in subsequent sections, we now introduce the interpolation spaces (power spaces) of RKHS. For any $s \geq 0$, the fractional power operator $L_k^s : L^2(\mathcal{X},\mu) \to L^2(\mathcal{X},\mu)$ is defined as

$$L_k^s(f) = \sum_{i\in N} \lambda_i^s \langle f, e_i\rangle_{L^2} e_i.$$

Then the interpolation space (power space) $[\mathcal{H}]^s$ is given by

$$[\mathcal{H}]^s := \operatorname{Ran} L_k^{s/2} = \left\{ \sum_{i\in N} a_i \lambda_i^{s/2} e_i : (a_i)_{i\in N} \in \ell_2(N) \right\} \subseteq L^2(\mathcal{X},\mu), \tag{3}$$

equipped with the inner product

$$\langle f, g\rangle_{[\mathcal{H}]^s} = \left\langle L_k^{-s/2} f, L_k^{-s/2} g \right\rangle_{L^2}.$$

It follows that $[\mathcal{H}]^s$ is also a separable Hilbert space with orthogonal basis $\{\lambda_i^{s/2} e_i\}_{i\in N}$. In particular, we have $[\mathcal{H}]^1 = \mathcal{H}$. For $0 < s_1 < s_2$, the embeddings $[\mathcal{H}]^{s_2} \hookrightarrow [\mathcal{H}]^{s_1} \hookrightarrow [\mathcal{H}]^0 \hookrightarrow L^2(\mathcal{X},\mu)$ are well-defined and compact (Fischer and Steinwart, 2020). For the functions in $[\mathcal{H}]^s$ with larger s, we say they have higher regularity (smoothness) with respect to the RKHS. In the remainder of this paper, we assume $|N| = \infty$. Also note that $\{\lambda_i\}_{i=1}^\infty$ and $\{e_i\}_{i=1}^\infty$ are dependent on $\mathcal{H}$, and hence depend on d.

3. Main results

3.1 KRR's generalization error in the general case

In this subsection, we consider a general framework for analyzing the generalization error of KRR, imposing only mild assumptions on the RKHS $\mathcal{H}$ and the true function f_ρ^*. In this

subsection, we allow d to grow with the sample size n and allow $\mathcal{H}$ and f_ρ^* to vary with d. Therefore, the results in this subsection are applicable to the large-dimensional setting $n \asymp d^\gamma, \gamma > 0$ (Section 3.2 and Section 3.3).

Given the RKHS $\mathcal{H}$ and the corresponding eigenvalues and eigenfunctions, the true function can be decomposed as $f_\rho^* = \sum_{i=1}^\infty f_i e_i(\boldsymbol{x}) \in L^2(\mathcal{X}, \mu)$ for some sequence $\{f_i\}_{i=1}^\infty$. We define the following important quantities, which are determined by $\{\lambda_i\}_{i=1}^\infty, \{e_i\}_{i=1}^\infty, \{f_i\}_{i=1}^\infty$ and regularization parameter λ:

$$\begin{aligned} &\mathcal{N}_1(\lambda) = \sum_{i=1}^\infty \left(\frac{\lambda_i}{\lambda_i + \lambda}\right); \quad \mathcal{N}_2(\lambda) = \sum_{i=1}^\infty \left(\frac{\lambda_i}{\lambda_i + \lambda}\right)^2; \\ &\mathcal{M}_1(\lambda) = \operatorname*{ess\,sup}_{\boldsymbol{x}\in\mathcal{X}} \left|\sum_{i=1}^\infty \left(\frac{\lambda}{\lambda_i + \lambda} f_i e_i(\boldsymbol{x})\right)\right|; \quad \mathcal{M}_2(\lambda) = \sum_{i=1}^\infty \left(\frac{\lambda}{\lambda_i + \lambda} f_i\right)^2. \end{aligned} \tag{4}$$

We emphasize that $\{\lambda_i\}_{i=1}^\infty, \{e_i\}_{i=1}^\infty, \{f_i\}_{i=1}^\infty$ depend on d, but the subscript d is omitted for brevity. Later in Theorem 1, we will express the result of generalization error through the quantities in (4).

Assumption 2 *Suppose that for some absolute constant $\sigma > 0$,*

$$\mathbb{E}_{(\boldsymbol{x},y)\sim\rho}\left[\left(y - f_\rho^*(\boldsymbol{x})\right)^2 \,\middle|\, \boldsymbol{x}\right] = \sigma^2, \quad \mu\text{-a.e. } \boldsymbol{x} \in \mathcal{X}.$$

Assumption 2 ensures that the noise is non-vanishing, a condition satisfied in the standard nonparametric regression model $y = f_\rho^*(\boldsymbol{x}) + \epsilon$ where ϵ is an independent nonzero noise.

Assumption 3 *Suppose that*

$$\operatorname*{ess\,sup}_{\boldsymbol{x}\in\mathcal{X}} \sum_{i=1}^\infty \left(\frac{\lambda_i}{\lambda_i + \lambda}\right)^2 e_i^2(\boldsymbol{x}) \le \mathcal{N}_2(\lambda), \tag{5}$$

and

$$\operatorname*{ess\,sup}_{\boldsymbol{x}\in\mathcal{X}} \sum_{i=1}^\infty \frac{\lambda_i}{\lambda_i + \lambda} e_i^2(\boldsymbol{x}) \le \mathcal{N}_1(\lambda). \tag{6}$$

Similar to Assumption 1, we suppose that Assumption 2 and 3 hold uniformly for all $d \ge 1$. Assumption 3 naturally holds for RKHSs with uniformly bounded eigenfunctions, i.e., $\sup_{i\ge1} \sup_{\boldsymbol{x}\in\mathcal{X}} |e_i(\boldsymbol{x})| \le 1$. Additionally, RKHSs associated with the inner product kernel on the unit sphere under the uniform distribution satisfy Assumption 3 (see Lemma 22).

Now we begin to state the first important theorem in this paper.

Theorem 1 *Let $\mathcal{N}_1, \mathcal{N}_2, \mathcal{M}_1$ and $\mathcal{M}_2$ be defined as (4), and let $d = d(n)$, which is allowed to grow as $n \to \infty$. Suppose that Assumption 1, 2 and 3 hold. Let $\hat{f}_\lambda$ be the KRR estimator defined by (1). If the following approximation conditions hold for some $\lambda = \lambda(d, n) \to 0$:*

$$\frac{\mathcal{N}_1(\lambda)}{n} \ln n = o(1); \quad n^{-1}\mathcal{N}_1(\lambda)^2 \ln n = o\left(\mathcal{N}_2(\lambda)\right); \quad n^{-1}\mathcal{N}_1(\lambda)^{\frac{1}{2}}\mathcal{M}_1(\lambda) = o\left(\mathcal{M}_2(\lambda)^{\frac{1}{2}}\right), \tag{7}$$

then we have

$$\mathbb{E}\left[\left\|\hat{f}_\lambda - f_\rho^*\right\|_{L^2}^2 \;\middle|\; \boldsymbol{X}\right] = \Theta_{\mathbb{P}}\left(\frac{\sigma^2\mathcal{N}_2(\lambda)}{n} + \mathcal{M}_2(\lambda)\right). \tag{8}$$

The notation $\Theta_{\mathbb{P}}$ only involves absolute constants.

Equation (8) presents the generalization error conditional on the input samples $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$, where the randomness in $\Theta_{\mathbb{P}}$ arises from $\boldsymbol{X}$. Theorem 1 provides the matching upper and lower bounds (8) for all λ satisfying the approximation conditions (7), where $\mathcal{N}_i(\lambda)$ and $\mathcal{M}_i(\lambda), i = 1, 2$, depend on d and n through $\lambda = \lambda(d, n)$. In (8), the term $\sigma^2\mathcal{N}_2(\lambda)/n$ corresponds to the variance, while $\mathcal{M}_2(\lambda)$ corresponds to the bias. Generally speaking, the conditions in (7) are more likely to hold for "larger" λ. For instance, we will show in the proof of Theorem 4 that if the conditions in (7) hold for $\lambda_0 = d^{-l_0}$, then they hold for all $\lambda = d^{-l}, 0 < l < l_0$.

Remark 2 *Theorem 1 shows that the generalization error is determined by $\mathcal{N}_1(\lambda), \mathcal{N}_2(\lambda)$, $\mathcal{M}_1(\lambda)$ and $\mathcal{M}_2(\lambda)$, and applies to both high-dimensional settings ($d = d(n) \to \infty$) and fixed-dimensional setting (d is fixed). Under the capacity condition (β) and source condition (s) framework (as discussed in Section 1) in the fixed-dimensional setting, Theorem 1 recovers the state-of-the-art results in Li et al. (2023b). Specifically, calculation shows that, as $\lambda = \lambda(n) \to 0$,*

$$\mathcal{N}_1(\lambda) \asymp \mathcal{N}_2(\lambda) \asymp \lambda^{-\frac{1}{\beta}}, \quad \mathcal{M}_1(\lambda) \lesssim \lambda^{\frac{\min\{s,2\}}{2}}, \quad \mathcal{M}_2(\lambda) \asymp \lambda^{\min\{s,2\}}, \tag{9}$$

where the bound of $\mathcal{M}_1(\lambda)$ actually requires an extra technical embedding index assumption (see, e.g., Li et al. 2023b). Therefore, Theorem 1 yields that, for $\lambda = \lambda(n) \gg n^{-\beta}$,

$$\mathbb{E}\left[\left\|\hat{f}_\lambda - f_\rho^*\right\|_{L^2}^2 \;\middle|\; \boldsymbol{X}\right] = \Theta_{\mathbb{P}}\left(\frac{\lambda^{-\frac{1}{\beta}}}{n} + \lambda^{\min\{s,2\}}\right),$$

which is consistent with Theorem 3.2 in Li et al. (2023b). As a corollary, choosing $\lambda \asymp n^{-\frac{\beta}{s\beta+1}}$ yields that KRR achieves the minimax optimal rate $n^{-\frac{s\beta}{s\beta+1}}$ when $0 < s \le 2$.

We emphasize that proving such tight bounds in the large-dimensional setting is nontrivial. In addition, the bounds in (9) under the capacity-source condition framework no longer hold in the large-dimensional setting. In the next subsection, we consider a specific setting, where exact convergence rates of $\mathcal{N}_1(\lambda), \mathcal{N}_2(\lambda), \mathcal{M}_1(\lambda)$ and $\mathcal{M}_2(\lambda)$ can be calculated, and derive concrete convergence rates through (8).

3.2 Applications to inner product kernel on the unit sphere

In this subsection, we consider the inner product kernel on the unit sphere $\mathbb{S}^d$ with uniform distribution. In the large-dimensional setting $n \asymp d^\gamma, \gamma > 0$, and under an additional source condition assumption, we apply Theorem 1 to establish the convergence rates of the generalization error of the KRR estimator. We then derive the corresponding minimax lower bound, allowing us to discuss the minimax optimality and the saturation effect of KRR.

Suppose that $\mathcal{X} = \mathbb{S}^d$ and μ is the uniform distribution on $\mathbb{S}^d$. We consider the inner product kernel, i.e., there exists a function $\Phi(t) : [-1, 1] \to \mathbb{R}$ such that $k_d(\boldsymbol{x}, \boldsymbol{x}') = \Phi\left(\langle \boldsymbol{x}, \boldsymbol{x}' \rangle\right), \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{S}^d$. Then Mercer's decomposition for the inner product kernel is then expressed in the basis of spherical harmonics:

$$k_d(\boldsymbol{x}, \boldsymbol{x}') = \sum_{k=0}^{\infty} \mu_k \sum_{l=1}^{N(d,k)} Y_{k,l}(\boldsymbol{x}) Y_{k,l}(\boldsymbol{x}'),$$

where $\{Y_{k,l}\}_{l=1}^{N(d,k)}$ are spherical harmonic polynomials of degree k; μ_k are the eigenvalues (also depending on d) with multiplicity $N(d, 0) = 1$; $N(d, k) = \frac{2k+d-1}{k} \frac{(k+d-2)!}{(d-1)!(k-1)!}, k = 1, 2, \cdots$.

Assumption 4 (Inner product kernel) *Suppose that $k = \{k_d\}_{d=1}^{\infty}$ satisfies*

$$k_d(\boldsymbol{x}, \boldsymbol{x}') = \Phi\left(\langle \boldsymbol{x}, \boldsymbol{x}' \rangle\right), \ \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{S}^d,$$

where $\Phi(t) \in \mathcal{C}^{\infty}([-1, 1])$ is a fixed function independent of d and

$$\Phi(t) = \sum_{j=0}^{\infty} a_j t^j, \ a_j > 0, \ \forall j = 0, 1, 2, \cdots$$

Assumption 4 formally defines the kernel considered in this subsection. We assume all coefficients $\{a_j\}_{j=0}^{\infty}$ are positive to simplify the main results and proofs. In fact, the proof remains similar for other inner product kernels, provided that the positive coefficients can be identified, for example, the neural tangent kernel in Section 3.3. We assume $\Phi(t)$ to be fixed (i.e., assume $\{a_j\}_{j=0}^{\infty}$ to be independent of d) and ignore the dependence of constants on $\{a_j\}_{j=0}^{\infty}$ in the rest of our paper.

The inner product kernel has attracted extensive research (Liang et al., 2020; Ghorbani et al., 2021; Misiakiewicz, 2022; Xiao et al., 2022; Lu et al., 2023, etc.). We have a concise characterization of μ_k and $N(d, k)$ for the inner product kernel on the unit sphere, which enables us to calculate the exact convergence rates of the key quantities in (4). We refer to Lemma 19, 20 and 21 in Appendix B.1 for details about μ_k and $N(d, k)$. The extension to general kernel can be extremely complicated and most existing results only consider the case where $\mathcal{X}$ is the sphere (as this paper) or discrete hypercube (see, e.g., Mei et al. 2022 and Aerni et al. 2022).

We next introduce the source condition, which characterizes the relative smoothness of f_ρ^* with respect to $\mathcal{H}$.

Assumption 5 (Source condition)

(a) Suppose that $f_\rho^*(\boldsymbol{x}) = f_{\rho_d}^*(\boldsymbol{x}) = \sum_{i=1}^{\infty} f_i e_i(\boldsymbol{x}) \in [\mathcal{H}]^s$ for some $s > 0$ and satisfies that,

$$\left\| f_\rho^* \right\|_{[\mathcal{H}]^s} \leq R_\gamma,$$

where R_γ is a constant only depending on γ.

(b) Denote q as the smallest integer such that $q > \gamma$ and $\mu_q \neq 0$. For $k \in \mathbb{N}$, define $\mathcal{I}_k$ as the index set satisfying $\lambda_i \equiv \mu_k, i \in \mathcal{I}_k$. Further suppose that there exists an absolute constant $c_0 > 0$ such that for any d and $k \in \{0, 1, \cdots, q\}$ with $\mu_k \neq 0$, we have

$$\sum_{i \in \mathcal{I}_k} \mu_k^{-s} f_i^2 \geq c_0. \tag{10}$$

Note that f_ρ^* and $\mathcal{H}$ vary with d, thus $\{f_i\}_{i=1}^\infty, \{\mu_k\}_{k=0}^\infty, \{\mathcal{I}_k\}_{k=0}^\infty$ also depend on d. We omit the subscript d and assume Assumption 5 to hold uniformly for all $d \geq 1$.

Remark 3 *Assumption 5 (a) is usually used as the traditional source condition (Caponnetto, 2006; Fischer and Steinwart, 2020, etc.). Assumption 5 (a) allows f_ρ^* to have a larger source condition, i.e., $f_\rho^* \in [\mathcal{H}]^{s'}$ for some $s' > s$. Consequently, we can only obtain the generalization error upper bound under Assumption 5 (a). Assumption 5 (b) is equivalent to assuming that the $[\mathcal{H}]^s$ norm of the projection of f_ρ^* on the first $(\lfloor \gamma \rfloor + 1)$-th eigenspaces (with multiplicity) is non-vanishing. Assumption 5 (b) enables us to study the worst generalization error convergence rate among functions f_ρ^* that belong precisely to $[\mathcal{H}]^s$. In other words, by calculating the interpolation norm (3), Assumption 5 implies: $f_\rho^* \in [\mathcal{H}]^s$ for any d; and for any $t > s$, $f_\rho^* \notin [\mathcal{H}]^t$ when d is sufficiently large.*

Now we are ready to state two theorems about the exact convergence rates of the generalization error of KRR, which deal with two different ranges of source condition: $s \geq 1$ and $0 < s < 1$.

Theorem 4 (Exact convergence rates when $s \geq 1$) *Suppose that there exist absolute constants $c_1, c_2, \gamma > 0$ such that, for any $d \in \mathbb{N}^+$, the sample size satisfies $c_1 d^\gamma \leq n \leq c_2 d^\gamma$. Consider $\mathcal{X} = \mathbb{S}^d$ and the marginal distribution μ_d to be the uniform distribution. Let $k = k_d$ be the inner product kernel on $\mathbb{S}^d$ satisfying Assumption 1 and 4. Further suppose that the true function $f_\rho^* = f_{\rho_d}^*$ satisfies Assumption 2 and Assumption 5 for some $s \geq 1$. Let $\hat{f}_\lambda$ be the KRR estimator defined by (1). Define $\tilde{s} = \min\{s, 2\}$, then we have:*

(i) If $\gamma \in (p + p\tilde{s},\ p + p\tilde{s} + 1]$ for some $p \in \mathbb{N}$, by choosing $\lambda = d^{-\frac{\gamma + p - p\tilde{s}}{2}} \cdot \mathbf{1}_{p>0} + d^{-\frac{\gamma}{2}} \ln d \cdot \mathbf{1}_{p=0}$, we have

$$\mathbb{E}\left[\left\| \hat{f}_\lambda - f_\rho^* \right\|_{L^2}^2 \;\middle|\; \boldsymbol{X} \right] = \begin{cases} \Theta_{\mathbb{P}}\left(d^{-\gamma} \ln^2 d\right) = \Theta_{\mathbb{P}}\left(n^{-1} \ln^2 n\right), & p = 0, \\ \Theta_{\mathbb{P}}\left(d^{-\gamma + p}\right) = \Theta_{\mathbb{P}}\left(n^{-1 + \frac{p}{\gamma}}\right), & p > 0; \end{cases}$$

(ii) If $\gamma \in (p + p\tilde{s} + 1,\ p + p\tilde{s} + 2\tilde{s} - 1]$ for some $p \in \mathbb{N}$, by choosing $\lambda = d^{-\frac{\gamma + 3p - p\tilde{s} + 1}{4}}$, we have

$$\mathbb{E}\left[\left\| \hat{f}_\lambda - f_\rho^* \right\|_{L^2}^2 \;\middle|\; \boldsymbol{X} \right] = \Theta_{\mathbb{P}}\left(d^{-\frac{\gamma - p + p\tilde{s} + 1}{2}}\right) = \Theta_{\mathbb{P}}\left(n^{-\frac{\gamma - p + p\tilde{s} + 1}{2\gamma}}\right);$$

(iii) If $\gamma \in (p + p\tilde{s} + 2\tilde{s} - 1,\ (p+1) + (p+1)\tilde{s}]$ for some $p \in \mathbb{N}$, by choosing $\lambda = d^{-\frac{\gamma + (p+1)(1-\tilde{s})}{2}}$, we have

$$\mathbb{E}\left[\left\| \hat{f}_\lambda - f_\rho^* \right\|_{L^2}^2 \;\middle|\; \boldsymbol{X} \right] = \Theta_{\mathbb{P}}\left(d^{-(p+1)\tilde{s}}\right) = \Theta_{\mathbb{P}}\left(n^{-\frac{(p+1)\tilde{s}}{\gamma}}\right).$$

The notation $\Theta_{\mathbb{P}}$ involves constants only depending on $s, \sigma, \gamma, c_0, \kappa, c_1$ and c_2. In addition, the convergence rates of the generalization error of KRR can not be faster than above for any choice of regularization parameter $\lambda = \lambda(d, n) \to 0$.

Theorem 5 (Exact convergence rates when $0 < s < 1$) *Suppose that there exist absolute constants $c_1, c_2, \gamma > 0$ such that, for any $d \in \mathbb{N}^+$, the sample size satisfies $c_1 d^\gamma \le n \le c_2 d^\gamma$. Consider $\mathcal{X} = \mathbb{S}^d$ and the marginal distribution μ_d to be the uniform distribution. Let $k = k_d$ be the inner product kernel on $\mathbb{S}^d$ satisfying Assumption 1 and 4. Further suppose that the true function $f_\rho^* = f_{\rho_d}^*$ satisfies Assumption 2 and Assumption 5 for some $0 < s < 1$. Let $\hat{f}_\lambda$ be the KRR estimator defined by (1). Then we have:*

- *If $\frac{1}{2} < s < 1$:*

  (i) If $\gamma \in (p + ps,\ p + ps + s]$ for some $p \in \mathbb{N}$, by choosing $\lambda = d^{-\frac{\gamma+p-ps}{2}} \cdot \mathbf{1}_{p>0} + d^{-\frac{\gamma}{2}} \ln d \cdot \mathbf{1}_{p=0}$, we have

$$\mathbb{E}\left[\left\|\hat{f}_\lambda - f_\rho^*\right\|_{L^2}^2 \;\middle|\; \boldsymbol{X}\right] = \begin{cases} \Theta_{\mathbb{P}}\left(d^{-\gamma} \ln^2 d\right) = \Theta_{\mathbb{P}}\left(n^{-1} \ln^2 n\right), & p = 0, \\ \Theta_{\mathbb{P}}\left(d^{-\gamma+p}\right) = \Theta_{\mathbb{P}}\left(n^{-1+\frac{p}{\gamma}}\right), & p > 0; \end{cases}$$

  (ii) If $\gamma \in (p + ps + s,\ (p+1) + (p+1)s]$ for some $p \in \mathbb{N}$, by choosing $\lambda = d^{-\frac{2p+s}{2}}$, we have

$$\mathbb{E}\left[\left\|\hat{f}_\lambda - f_\rho^*\right\|_{L^2}^2 \;\middle|\; \boldsymbol{X}\right] = \Theta_{\mathbb{P}}\left(d^{-(p+1)s}\right) = \Theta_{\mathbb{P}}\left(n^{-\frac{(p+1)s}{\gamma}}\right);$$

  The notation $\Theta_{\mathbb{P}}$ involves constants only depending on $s, \sigma, \gamma, c_0, \kappa, c_1$ and c_2.

- *If $0 < s \le \frac{1}{2}$: we have the same convergence rates as the case $s \in (\frac{1}{2}, 1)$ for those*

$$\gamma > \frac{3s}{2(s+1)}.$$

Remark 6 *For technical reasons, when $0 < s \le 1/2$, we only prove the convergence rates for those $\gamma > 3s/2(s+1)$. Note that we have $3s/2(s+1) < 1/2$ when $0 < s \le 1/2$; and $3s/2(s+1) \to 0$ when $s \to 0$. Therefore, we have actually proved for almost all $\gamma > 0$.*

Theorem 4 and Theorem 5 establish exact convergence rates (both upper and lower bounds) of KRR's generalization error, which is a significantly stronger result than proving only an upper bound. As we will see in Appendix B.4, since $\|f_\rho^*\|_{L^\infty}$ could be infinite when $s < 1$ thus $\mathcal{M}_1(\lambda)$ could be infinite, the proof of Theorem 5 requires a little more technique. In addition, we will prove in Theorem 7 that the rates in Theorem 5 ($s \le 1$) achieve the minimax lower bound. Together with the statement at the end of Theorem 4, we actually prove that the rates in Theorem 4 and Theorem 5 are the fastest convergence rates that KRR can achieve.

Theorem 4 and Theorem 5 reveal that the convergence rate varying along γ exhibits periodicity, specifically $\gamma \in \big(p + ps,\ (p+1) + (p+1)s\big]$ for each $p \in \mathbb{N}$. This is due to the special form of eigenspaces of the inner product kernel. Specifically, the properties of the

eigenvalues (Lemma 19 and Lemma 21) yield that: by choosing $\lambda = d^l, l > 0$, the order of $\mathcal{N}_2(\lambda)$ and $\mathcal{M}_2(\lambda)$ in Theorem 1 has the period $l \in (p, p+1]$ for each $p \in \mathbb{N}$ (Lemma 23 and Lemma 25). Balancing the two term in the generalization error bound (8), i.e., $\mathcal{N}_2(\lambda)/n$ and $\mathcal{M}_2(\lambda)$, yields specific relation between l and γ, which finally yields the period of $\gamma \in \big(p+ps,\ (p+1)+(p+1)s\big]$. We refer to Appendix B.3 and Appendix B.4 for detailed calculations.

Next, we will state the minimax lower bound in the same large-dimensional and source condition setting as Theorem 4 and Theorem 5.

Theorem 7 (Minimax lower bound) *Suppose that there exist absolute constants $c_1, c_2, \gamma > 0$ such that, for any $d \in \mathbb{N}^+$, the sample size satisfies $c_1 d^\gamma \le n \le c_2 d^\gamma$. Consider $\mathcal{X} = \mathbb{S}^d$ and the marginal distribution μ_d to be the uniform distribution. Let $k = k_d$ be the inner product kernel on $\mathbb{S}^d$ satisfying Assumption 1 and 4. Let $\mathcal{P} = \mathcal{P}_d$ consists of all the distributions $\rho = \rho_d$ on $\mathcal{X} \times \mathcal{Y}$ such that Assumption 2 holds and Assumption 5 holds for some $s > 0$. Then we have:*

(i) If $\gamma \in (p+ps, p+ps+s]$ for some $p \in \mathbb{N}$, for any $\varepsilon > 0$, there exist constants C_1 and C only depending on $s, \varepsilon, \gamma, \sigma, \kappa, c_1$ and c_2 such that for any $d \ge C$, we have:

$$\min_{\hat{f}} \max_{\rho \in \mathcal{P}} \mathbb{E}_{(\boldsymbol{X},\mathbf{y}) \sim \rho^{\otimes n}} \left\| \hat{f} - f_\rho^* \right\|_{L^2}^2 \ge C_1 d^{-\gamma + p - \varepsilon}; \tag{11}$$

(ii) If $\gamma \in (p+ps+s, (p+1)+(p+1)s]$ for some $p \in \mathbb{N}$, there exist constants C_1 and C only depending on $s, \gamma, \sigma, \kappa, c_1$ and c_2 such that for any $d \ge C$, we have:

$$\min_{\hat{f}} \max_{\rho \in \mathcal{P}} \mathbb{E}_{(\boldsymbol{X},\mathbf{y}) \sim \rho^{\otimes n}} \left\| \hat{f} - f_\rho^* \right\|_{L^2}^2 \ge C_1 d^{-(p+1)s}. \tag{12}$$

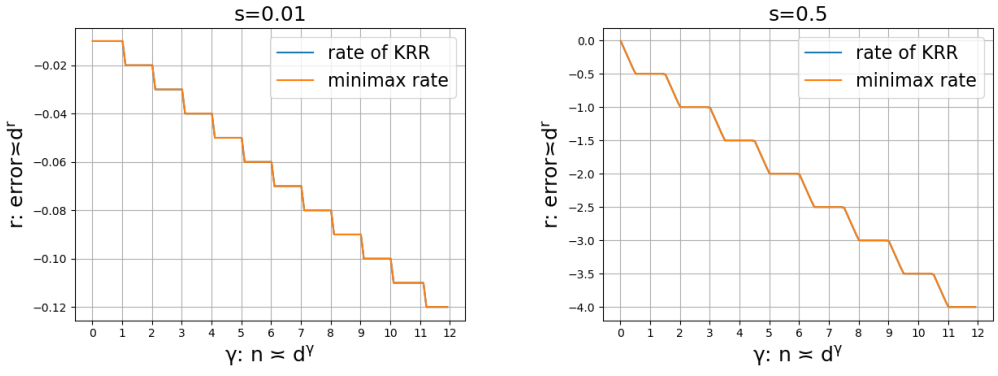
Theorem 7 states that no estimator (or learning method) that can achieve faster convergence rates than those given in (11) and (12).

Figure 2 summarizes the results in Theorem 4, Theorem 5 and Theorem 7, illustrating the convergence rates of KRR and the corresponding minimax lower rates *with respect to dimension d* for any $\gamma > 0$. We can see that the rates decrease when the scaling γ increases, indicating that the performance becomes better when the sample size n grows. Moreover, we can observe several intriguing phenomena.

Curve's evolution with source condition. Since we consider source condition $s > 0$, we can compare the rate curves in Figure 2 for different s and see how they evolve with s.

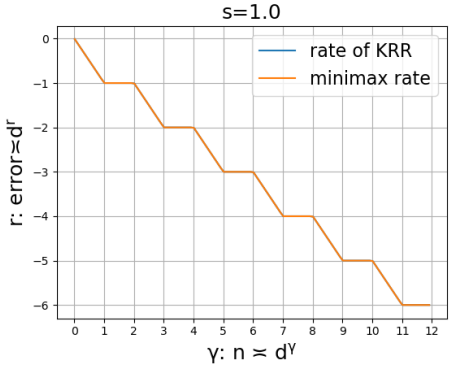
Let us first see the minimax lower rates. For any $s > 0$, there are 2 periods with respect to the value of γ: The first period, $(p+ps, p+ps+s], p \in \mathbb{N}$, shrinks to 0 as s approaches 0; The second period, $(p+ps+s, (p+1)+(p+1)s]$, remains constant at length 1 for all $s > 0$. Next, we examine the convergence rates of KRR, which exhibit more intricate behavior.

- When $0 < s \le 1$, there are 2 periods with respect to the value of γ and the curve is the same as the minimax lower rates. (In fact, Theorem 5 only proves the results for $\gamma > 3s/2(s+1)$ when $s \le 1/2$, we write $\gamma > 0$ with a little bit of notation abusement.)

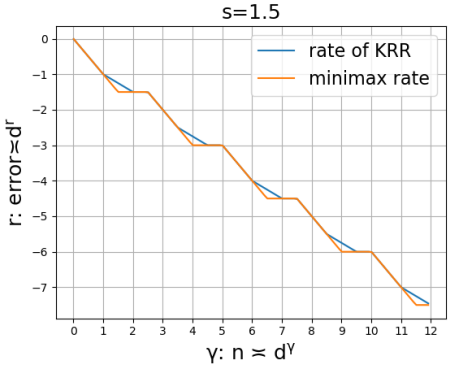


(a)

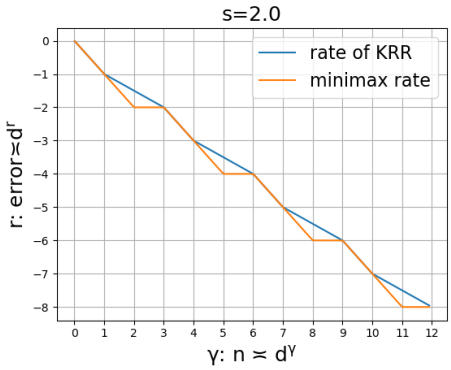
(b)



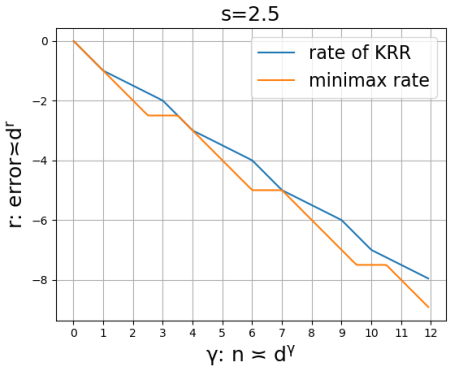
(c)



(d)



(e)



(f)

Figure 2: Convergence rates of KRR in Theorem 4, Theorem 5 and corresponding minimax lower rates in Theorem 7 (ignoring a ε-difference) *with respect to dimension* d. We present six graphs corresponding to six different source conditions: $s = 0.01, 0.5, 1.0, 1.5, 2.0, 2.5$. The x-axis represents asymptotic scaling, $\gamma : n \asymp d^{\gamma}$; the y-axis represents the convergence rate of generalization error, r : error $\asymp d^{r}$.

- When $1 < s < 2$, there are 3 periods with respect to the value of γ: The length of the first period, i.e., $(p + ps, p + ps + 1]$, equals 1 for all $1 < s < 2$; The length of the second period, i.e., $(p + ps + 1, p + ps + 2s - 1]$ is $2s - 2$, thus this period will degenerate as s getting close to 1; The length of the third period, i.e., $(p + ps + 2s - 1, (p + 1) + (p + 1)s]$ is $2 - s$, thus this period will degenerate as s getting close to 2.

- When $s \geq 2$, the curve does not change with s and there are 2 periods with respect to the value of γ: The length of the first period, i.e., $(3p, 3p + 1]$, equals 1 for all $s \geq 2$; The length of the second period, i.e., $(3p + 1, 3p + 3]$, equals 2 for all $s \geq 2$.

Minimax optimality and new saturation effect of KRR. Figure 2 (a)(b)(c) show that the convergence rates of KRR match the minimax lower bound for all $\gamma > 0$, establishing the minimax optimality of KRR when $0 < s \leq 1$. In contrast, when $s > 1$, Figure 2 (d)(e)(f) illustrate that KRR can not achieve the minimax lower bound in Theorem 7 for certain ranges of γ, which we refer to as the "new saturation effect" of KRR.

In the fixed-dimensional setting, the saturation effect (Li et al., 2023a) says that when source condition satisfies $s > 2$, no matter how carefully KRR is tuned, the convergence rate can not achieve the minimax lower bound. Specifically, when the capacity condition (β) and source condition (s) are assumed in the fixed-dimensional setting, the best convergence rate of KRR is $n^{-\frac{2\beta}{2\beta+1}}$ when $s > 2$, which does not achieve the minimax lower bound $n^{-\frac{s\beta}{s\beta+1}}$. In the large-dimensional setting and for inner product kernel on the unit sphere, our results show that the saturation effect of KRR happens in a new regime $1 < s \leq 2$. In addition, we conjecture that there are other spectral algorithms (e.g., kernel gradient flow) that can achieve the minimax lower bound in Theorem 7 for all $s > 0$.

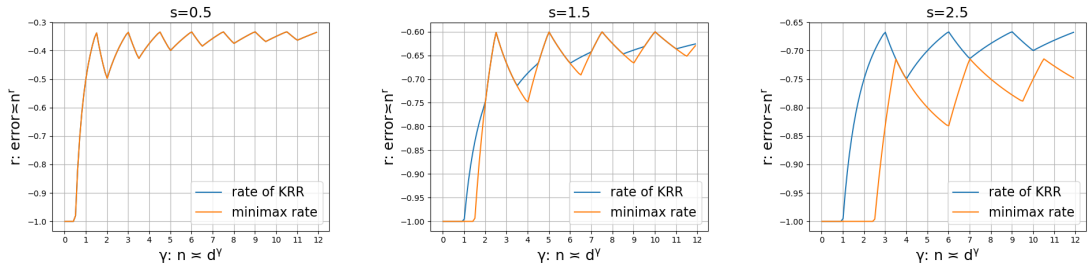
Periodic plateau behavior. If $0 < s < 2$, Figure 2 (a)(b)(c) show that within specific intervals of γ, the vertical axis value, r, remains constant. We refer to such ranges of γ as the plateau period. When s exceeds 2, the plateau period of KRR's convergence rates degenerates and the plateau period of minimax lower rates still exists. Also note that the length of each plateau period varies with the values $s > 0$.

For these plateau periods, if we fix a large dimension d and increase γ (or equivalently, increase the sample size n), the convergence rates of KRR or minimax lower rates stay invariant in certain ranges. Therefore, in order to improve the rate, one has to increase the sample size above a certain threshold.

Figure 3 provides an alternative representation of our results, which shows the convergence rates of KRR and corresponding minimax lower rates *with respect to sample size n*. We can observe the "multiple descent behavior" (for both the convergence rates of KRR and the minimax lower rates) from Figure 3.

Multiple descent behavior. We first examine the minimax lower rates. For any $s > 0$, the curve attains peaks at $\gamma = p + ps, p \in \mathbb{N}^+$, and reaches isolated valleys at $\gamma = p + ps + s, p \in \mathbb{N}^+$. Next, we examine the convergence rates of KRR:

- When $0 < s \leq 1$, the curve is the same as the curve of minimax lower rates.



(a) (b) (c)

Figure 3: Convergence rates of KRR in Theorem 4, Theorem 5 and corresponding minimax lower rates in Theorem 7 (ignoring a ε-difference) *with respect to sample size* n. We present 3 graphs corresponding to 3 kinds of source conditions: $s = 0.5, 1.5, 2.5$. The x-axis represents asymptotic scaling, $\gamma : n \asymp d^{\gamma}$; the y-axis represents the convergence rate of generalization error, $r : \text{error} \asymp n^{r}$.

- When $1 < s < 2$, the curve achieves its peaks at $\gamma = p + p\tilde{s}, p \in \mathbb{N}^+$; achieve its isolated valleys at $\gamma = p + p\tilde{s} + 1, p \in \mathbb{N}^+$ and achieve its hillside at $\gamma = p + p\tilde{s} + 2\tilde{s} - 1, p \in \mathbb{N}^+$.
- When $s \geq 2$, the curve does not change with s, which achieves its peaks at $\gamma = 3p, p \in \mathbb{N}^+$, and achieve its isolated valleys at $\gamma = 3p + 1, p \in \mathbb{N}^+$.

3.3 Applications to neural tangent kernel

In this subsection, we consider a specific example: the neural tangent kernel (NTK) of a two-layer fully connected ReLU neural network $k_d = k_d^{\mathrm{NT}}$. We continue to suppose that $\mathcal{X} = \mathbb{S}^d$ and μ is the uniform distribution on $\mathbb{S}^d$. It has been shown in Bietti and Mairal (2019); Lu et al. (2023) that the NTK is an example of inner product kernel satisfying

$$k_d^{\mathrm{NT}}(\boldsymbol{x}, \boldsymbol{x}') = \Phi\left(\langle \boldsymbol{x}, \boldsymbol{x}' \rangle\right), \ \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{S}^d,$$

where $\Phi(t) = \sum_{j=0}^{\infty} a_j t^j \in \mathcal{C}^{\infty}([-1, 1])$ is a fixed function independent of d and

$$a_0 > 0; \quad a_1 > 0; \quad a_j > 0, \ \forall j = 2, 4, 6, \cdots; \quad a_j = 0, \ \forall j = 3, 5, 7, \cdots$$

Lemma 5 in Lu et al. (2023) also shows that $\sup_{\boldsymbol{x} \in \mathcal{X}} k_d^{\mathrm{NT}}(\boldsymbol{x}, \boldsymbol{x}) \leq 1$.

For the neural tangent kernel k_d^{NT}, the following theorems establish the exact convergence rates of the generalization error of KRR and the corresponding minimax lower bound. As the proofs are similar to Theorem 4, Theorem 5 and Theorem 7, we omit the proofs of the following theorems for brevity. Throughout this subsection, we define $\mathcal{I}_{\mathrm{NT}} = \{0, 1\} \cup \{2, 4, 6, \cdots\}$, where $a_j > 0$ for all $j \in \mathcal{I}_{\mathrm{NT}}$.

Theorem 8 (NTK: exact convergence rates when s ≥ 1) *Suppose that there exist absolute constants* $c_1, c_2, \gamma > 0$ *such that, for any* $d \in \mathbb{N}^+$, *the sample size satisfies* $c_1 d^{\gamma} \leq n \leq c_2 d^{\gamma}$. *Consider* $\mathcal{X} = \mathbb{S}^d$ *and the marginal distribution* μ_d *to be the uniform distribution. Let*

$k = k_d^{\mathrm{NT}}$ be the neural tangent kernel of a two-layer fully connected ReLU neural network on $\mathbb{S}^d$. Further suppose that the true function $f_\rho^* = f_{\rho_d}^*$ satisfies Assumption 2 and Assumption 5 for some $s \geq 1$. Let $\hat{f}_\lambda$ be the KRR estimator defined by (1). For any $p \in \mathcal{I}_{\mathrm{NT}}$, we define $p' = p + 2$, if $p \geq 2$; and $p' = p + 1$, if $p \leq 1$. Define $\tilde{s} = \min\{s, 2\}$, then we have:

(i) If $\gamma \in (p + p\tilde{s}, p' + p\tilde{s}]$ for some $p \in \mathcal{I}_{\mathrm{NT}}$, by choosing $\lambda = d^{-\frac{\gamma+p-p\tilde{s}}{2}} \cdot \mathbf{1}_{p>0} + d^{-\frac{\gamma}{2}} \ln d \cdot \mathbf{1}_{p=0}$, we have

$$\mathbb{E}\left[\left\|\hat{f}_\lambda - f_\rho^*\right\|_{L^2}^2 \;\middle|\; \boldsymbol{X}\right] = \begin{cases} \Theta_{\mathbb{P}}\left(d^{-\gamma} \ln^2 d\right) = \Theta_{\mathbb{P}}\left(n^{-1} \ln^2 n\right), & p = 0, \\ \Theta_{\mathbb{P}}\left(d^{-\gamma+p}\right) = \Theta_{\mathbb{P}}\left(n^{-1+\frac{p}{\gamma}}\right), & p > 0; \end{cases}$$

(ii) If $\gamma \in (p' + p\tilde{s}, 2p'\tilde{s} - p' + 2p - p\tilde{s}]$ for some $p \in \mathcal{I}_{\mathrm{NT}}$, by choosing $\lambda = d^{-\frac{\gamma+p'+2p-p\tilde{s}}{4}}$, we have

$$\mathbb{E}\left[\left\|\hat{f}_\lambda - f_\rho^*\right\|_{L^2}^2 \;\middle|\; \boldsymbol{X}\right] = \Theta_{\mathbb{P}}\left(d^{-\frac{\gamma}{2}+\frac{p'}{2}-\frac{p\tilde{s}}{2}-2}\right) = \Theta_{\mathbb{P}}\left(n^{-\frac{1}{2}+\frac{p'}{2\gamma}-\frac{p\tilde{s}}{2\gamma}-\frac{2}{\gamma}}\right);$$

(iii) If $\gamma \in (2p'\tilde{s} - p' + 2p - p\tilde{s}, p' + p'\tilde{s}]$ for some $p \in \mathcal{I}_{\mathrm{NT}}$, by choosing $\lambda = d^{-\frac{\gamma+p'(1-\tilde{s})}{2}}$, we have

$$\mathbb{E}\left[\left\|\hat{f}_\lambda - f_\rho^*\right\|_{L^2}^2 \;\middle|\; \boldsymbol{X}\right] = \Theta_{\mathbb{P}}\left(d^{-p'\tilde{s}}\right) = \Theta_{\mathbb{P}}\left(n^{-\frac{p'\tilde{s}}{\gamma}}\right).$$

The notation $\Theta_{\mathbb{P}}$ involves constants only depending on $s, \sigma, \gamma, c_0, c_1$ and c_2.

Theorem 9 (NTK: exact convergence rates when $0 < s < 1$) *Suppose that there exist absolute constants $c_1, c_2, \gamma > 0$ such that, for any $d \in \mathbb{N}^+$, the sample size satisfies $c_1 d^\gamma \leq n \leq c_2 d^\gamma$. Consider $\mathcal{X} = \mathbb{S}^d$ and the marginal distribution μ_d to be the uniform distribution. Let $k = k_d^{\mathrm{NT}}$ be the neural tangent kernel of a two-layer fully connected ReLU neural network on $\mathbb{S}^d$. Further suppose that the true function $f_\rho^* = f_{\rho_d}^*$ satisfies Assumption 2 and Assumption 5 for some $0 < s < 1$. Let $\hat{f}_\lambda$ be the KRR estimator defined by (1). For any $p \in \mathcal{I}_{\mathrm{NT}}$, we define $p' = p + 2$, if $p \geq 2$; and $p' = p + 1$, if $p \leq 1$. Then we have:*

- *If $\frac{1}{2} < s < 1$:*

  (i) When $\gamma \in (p + ps, p + p's]$ for some $p \in \mathcal{I}_{\mathrm{NT}}$, by choosing $\lambda = d^{-\frac{\gamma+p-ps}{2}} \cdot \mathbf{1}_{p>0} + d^{-\frac{\gamma}{2}} \ln d \cdot \mathbf{1}_{p=0}$, we have

$$\mathbb{E}\left[\left\|\hat{f}_\lambda - f_\rho^*\right\|_{L^2}^2 \;\middle|\; \boldsymbol{X}\right] = \begin{cases} \Theta_{\mathbb{P}}\left(d^{-\gamma} \ln^2 d\right) = \Theta_{\mathbb{P}}\left(n^{-1} \ln^2 n\right), & p = 0, \\ \Theta_{\mathbb{P}}\left(d^{-\gamma+p}\right) = \Theta_{\mathbb{P}}\left(n^{-1+\frac{p}{\gamma}}\right), & p > 0; \end{cases}$$

  (ii) When $\gamma \in (p + p's, p' + p's]$ for some $p \in \mathcal{I}_{\mathrm{NT}}$, by choosing $\lambda = d^{-p-\frac{(p'-p)s}{2}}$, we have

$$\mathbb{E}\left[\left\|\hat{f}_\lambda - f_\rho^*\right\|_{L^2}^2 \;\middle|\; \boldsymbol{X}\right] = \Theta_{\mathbb{P}}\left(d^{-p's}\right) = \Theta_{\mathbb{P}}\left(n^{-\frac{p's}{\gamma}}\right);$$

  The notation $\Theta_{\mathbb{P}}$ involves constants only depending on $s, \sigma, \gamma, c_0, c_1$ and c_2.

- *If $0 < s \le \frac{1}{2}$: we have the same convergence rates as the case $s \in (\frac{1}{2}, 1)$ for those*

$$\gamma > \frac{3s}{2(s+1)}.$$

Theorem 10 (NTK: minimax lower bound) *Suppose that there exist absolute constants $c_1, c_2, \gamma > 0$ such that, for any $d \in \mathbb{N}^+$, the sample size satisfies $c_1 d^\gamma \le n \le c_2 d^\gamma$. Consider $\mathcal{X} = \mathbb{S}^d$ and the marginal distribution μ_d to be the uniform distribution. Let $k = k_d$ be the neural tangent kernel of a two-layer fully connected ReLU neural network on $\mathbb{S}^d$. Let $\mathcal{P} = \mathcal{P}_d$ denote the set of all the distributions $\rho = \rho_d$ on $\mathcal{X} \times \mathcal{Y}$ such that Assumption 2 holds and Assumption 5 holds for some $s > 0$. Then we have:*

(i) When $\gamma \in (p + ps, p + p's]$ for some $p \in \mathcal{I}_{\mathrm{NT}}$, for any $\varepsilon > 0$, there exist constants C_1 and C only depending on $s, \varepsilon, \gamma, \sigma, c_1$ and c_2 such that for any $d \ge C$, we have:

$$\min_{\hat{f}} \max_{\rho \in \mathcal{P}} \mathbb{E}_{(\boldsymbol{X}, \mathbf{y}) \sim \rho^{\otimes n}} \left\| \hat{f} - f_\rho^* \right\|_{L^2}^2 \ge C_1 d^{-\gamma + p - \varepsilon};$$

(ii) When $\gamma \in (p + p's, p' + p's]$ for some $p \in \mathcal{I}_{\mathrm{NT}}$, there exist constants C_1 and C only depending on s, γ, σ, c_1 and c_2 such that for any $d \ge C$, we have:

$$\min_{\hat{f}} \max_{\rho \in \mathcal{P}} \mathbb{E}_{(\boldsymbol{X}, \mathbf{y}) \sim \rho^{\otimes n}} \left\| \hat{f} - f_\rho^* \right\|_{L^2}^2 \ge C_1 d^{-p's};$$

The results in this subsection are consistent with the theorems in Section 3.2 if we redefine $p' = p + 1, \forall p = 0, 1, 2, \cdots$.

4. Conclusion and discussion

In this paper, we first establish a new framework for studying the asymptotic generalization error of kernel ridge regression (Theorem 1). This framework imposes minimal assumptions on the RKHS, the true function and the relation between d and n, making it suitable for studying various topics about KRR's generalization error in both fixed-dimensional and large-dimensional settings. Moreover, the Theorem 1 establishes the matching upper and lower bounds of the generalization error with a suitable choice of the regularization parameter, which is more informative than just the upper bound.

Building on this framework, we analyze the inner product kernel on the unit sphere and the large-dimensional setting ($n \asymp d^\gamma, \gamma > 0$). Assuming that f_ρ^* belongs to $[\mathcal{H}]^s$, an interpolation space of the RKHS, Theorem 4 and Theorem 5 establish the exact convergence rates of KRR's generalization error under the best choice of regularization parameter and Theorem 7 proves the corresponding minimax lower bound. These results show the minimax optimality of KRR when $0 < s \le 1$ and the new saturation effect of KRR $s > 1$. We also discuss how the convergence rate curves evolve with the value of s and highlight the "periodic plateau behavior" (for the rate w.r.t. d) and "multiple descent behavior" (for the rate w.r.t. n) in the large-dimensional setting.

Similar periodic behavior has been observed in prior studies on kernel methods in the large-dimensional setting. We now discuss key related works. There is a line of work studying

the inconsistency of kernel methods with inner product kernels in the large-dimensional setting $n \asymp d^\gamma, \gamma > 0$ (Ghorbani et al., 2021; Mei et al., 2022; Misiakiewicz, 2022, etc.). Assuming the true function f_ρ^* to be square-integrable (or equivalently $s = 0$ in our setting) on the unit sphere, Ghorbani et al. (2021, Theorem 4) proves that the generalization error of KRR $R_{\mathrm{KR}}\left(f_\rho^*, \boldsymbol{X}, \lambda\right)$ satisfies (with high probability)

$$\left|R_{\mathrm{KR}}\left(f_\rho^*, \boldsymbol{X}, \lambda\right) - \left\|\mathrm{P}_{>\ell} f_\rho^*\right\|_{L^2}^2\right| \leq \varepsilon\left(\left\|f_\rho^*\right\|_{L^2}^2 + \sigma^2\right), \forall\ 0 < \lambda < \lambda^*, \tag{13}$$

where $\ell = \lfloor \gamma \rfloor$ is the greatest integer that is less or equal to γ, $\mathrm{P}_{>\ell}$ means the projection onto polynomials with degree $> \ell$, ε is any positive real number and λ^* is defined as Ghorbani et al. 2021, Eq.(20). (13) implies that generalization error drops sharply when γ crosses an integer and remains constant otherwise (see the schematic illustration in Ghorbani et al. 2021, Figure 5). In our paper, when $s > 0$ and sufficiently close to 0, similar behavior can be observed in Figure 2 (a) that the rate drops abruptly around each integer $\gamma \in \mathbb{N}$.

A more recent work Lu et al. (2023) considers the optimality of early stopping kernel gradient flow in the same large-dimensional setting $c_1 d^\gamma \leq n \leq c_2 d^\gamma, \gamma > 0$. They also consider inner product kernel on the sphere and assume that the true function falls into the RKHS $f_\rho^* \in \mathcal{H}$ (or equivalently, $s = 1$). Denoting $p = \lfloor \gamma/2 \rfloor$, Lu et al. 2023, Theorem 4.3 proves that by properly choosing the early stopping time $\widehat{T}$, the upper bound of the convergence rate is:

- When $\gamma \in \{2, 4, 6, \cdots\}$, then, there exist constants C and C_i, where $i = 1, 2, 3$, only depending on γ, c_1, and c_2, such that for any $d \geq C$, we have

  $$\left\|f_{\widehat{T}} - f_\star\right\|_{L^2}^2 \leq C_1 n^{-\frac{1}{2}}$$

  holds with probability at least $1 - C_2 \exp\{-C_3 n^{1/2}\}$.

- When $\gamma \in \bigcup_{j=0}^{\infty}(2j, 2j+1]$, for any $\delta > 0$, there exist constants C and C_i, where $i = 1, 2, 3$, only depending on γ, δ, c_1, and c_2, such that for any $d \geq C$, we have

  $$\left\|f_{\widehat{T}} - f_\rho^*\right\|_{L^2}^2 \leq C_1 n^{-\frac{\gamma - p}{\gamma}} \log(n)$$

  holds with probability at least $1 - \delta - C_2 \exp\{-C_3 n^{p/\gamma} \log(n)\}$.

- When $\gamma \in \bigcup_{j=0}^{\infty}(2j+1, 2j+2)$, then, for any $\delta > 0$, there exist constants C and C_i, where $i = 1, 2, 3$, only depending on γ, δ, c_1, and c_2, such that for any $d \geq C$, we have

  $$\left\|f_{\widehat{T}} - f_\star\right\|_{L^2}^2 \leq C_1 n^{-\frac{p+1}{\gamma}}$$

  holds with probability at least $1 - \delta - C_2 \exp\{-C_3 n^{1-(p+1)/\gamma}\}$.

Ignoring the logarithmic factor, a straightforward calculation reveals that the convergence rate coincides with the rate in Theorem 4 when $s = 1$ (see Figure 2 (c)). Lu et al. (2023) also proves that the above upper bound matches the minimax lower bound, thus proving the minimax optimality of early stopping kernel gradient flow under the assumption $f_\rho^* \in \mathcal{H}$. In contrast to Lu et al. (2023), we provide both the upper bound and the lower bound of

the convergence rates of KRR under general source condition $s > 0$. We have seen that the "periodic plateau behavior" (for the rate w.r.t. d) and "multiple descent behavior" (for the rate w.r.t. n) observed in Lu et al. (2023) still exist for $s > 0$ and the plateau length will change with the value of s. By incorporating source condition $s > 0$, we relax constraints on the true function, offering a more complete characterization of the generalization error of KRR.

Periodic behavior in large dimension has also been observed for "kernel interpolation estimator", for instance, Liang et al. (2020) for the inner product kernel and Aerni et al. (2022) for the convolutional kernel. While technically challenging, a direct follow-up question is the convergence rate of generalization error for general kernels and domains. We believe that it is an interesting research direction to study the generalization behavior of kernel methods in the large-dimensional setting, which will exhibit a wealth of new phenomena compared with the fixed-dimensional setting. In addition, finding optimal $\lambda(d, n)$ (even in the fixed-dimensional setting) is difficult in practice. Cross-validation is a widely used method to select optimal λ adaptively, and it has been shown to be effective for several kernel regression algorithms (Caponnetto and Yao, 2010; Raskutti et al., 2014). However, to our knowledge, this issue remains unexplored in the large-dimensional setting. We believe that developing adaptive estimators is another meaningful future question.

Acknowledgments

The authors are grateful to the editors and the referees for their constructive comments that greatly improved the quality and the presentation of this paper. Qian Lin is supported in part by the National Natural Science Foundation of China (Grant 92370122, Grant 11971257) and the Beijing Natural Science Foundation (Grant Z190001).

Appendix

In the appendix, we provide the proof of Theorem 1 (Appendix A), Theorem 4 & 5 (Appendix B) and Theorem 7 (Appendix C). Appendix D contains the auxiliary results.

Appendix A. Proof of Theorem 1

The proof of Theorem 1 consists of the following steps: First, we introduce the bias-variance decomposition in Section A.1. Next, we derive the bounds of variance term in Section A.2 and bias term in Section A.3. Finally, using the results in these sections, we formally prove Theorem 1 in Section A.4.

A.1 Bias-variance decomposition

The proof of Theorem 1 is based on the traditional bias-variance decomposition. The contribution in this paper is that we refine the tools in Li et al. (2024) and Li et al. (2023b) to extend the applicability to the large-dimensional case. Throughout the proof, we denote

$$T_\lambda = T + \lambda; \quad T_{\boldsymbol{X}\lambda} = T_{\boldsymbol{X}} + \lambda, \tag{14}$$

where λ is the regularization parameter. $T+\lambda$ actually means $T+\lambda I$, where I is the identity operator. We use $\|\cdot\|_{\mathscr{B}(B_1,B_2)}$ to denote the operator norm of a bounded linear operator from a Banach space B_1 to B_2, i.e., $\|A\|_{\mathscr{B}(B_1,B_2)} = \sup_{\|f\|_{B_1}=1} \|Af\|_{B_2}$. For brevity, we denote it simply as $\|\cdot\|$ where no ambiguity arises. In addition, we use $\mathrm{tr}A$ and $\|A\|_1$ to denote the trace and the trace norm of an operator. We denote $\|A\|_2$ as the Hilbert-Schmidt norm. In addition, we use $L^2(\mathcal{X},\mu)$ and $L^\infty(\mathcal{X},\mu)$ simply as L^2 and L^∞ throughout the proof for brevity.

We also need the following essential notations, which are frequently used in related literature. Denote the samples $\boldsymbol{Z} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$. Define the sampling operator $K_{\boldsymbol{x}} : \mathbb{R} \to \mathcal{H},\ y \mapsto yk(\boldsymbol{x},\cdot)$ and its adjoint operator $K_{\boldsymbol{x}}^* : \mathcal{H} \to \mathbb{R},\ f \mapsto f(\boldsymbol{x})$. Then we can define $T_{\boldsymbol{x}} = K_{\boldsymbol{x}} K_{\boldsymbol{x}}^*$. Furthermore, we define the sample covariance operator $T_{\boldsymbol{X}} : \mathcal{H} \to \mathcal{H}$ as

$$T_{\boldsymbol{X}} := \frac{1}{n}\sum_{i=1}^n K_{\boldsymbol{x}_i} K_{\boldsymbol{x}_i}^*. \tag{15}$$

Then we know that $\|T_{\boldsymbol{X}}\| \le \|T_{\boldsymbol{X}}\|_1 \le \kappa^2$ and $T_{\boldsymbol{X}}$ is a trace class thus compact operator. Furthermore, define the sample basis function

$$g_{\boldsymbol{Z}} := \frac{1}{n}\sum_{i=1}^n K_{\boldsymbol{x}_i} y_i \in \mathcal{H}.$$

Following Caponnetto and de Vito (2007), the KRR estimator (1) can be expressed in operator form as

$$\hat{f}_\lambda = (T_{\boldsymbol{X}} + \lambda)^{-1} g_{\boldsymbol{Z}},$$

In order to derive the bias term, we define

$$\tilde{g}_{\boldsymbol{Z}} := \mathbb{E}\left(g_{\boldsymbol{Z}} | \boldsymbol{X}\right) = \frac{1}{n}\sum_{i=1}^n K_{\boldsymbol{x}_i} f_\rho^*(\boldsymbol{x}_i) \in \mathcal{H};$$

and

$$\tilde{f}_\lambda := \mathbb{E}\left(\hat{f}_\lambda | \boldsymbol{X}\right) = \left(T_{\boldsymbol{X}} + \lambda\right)^{-1} \tilde{g}_{\boldsymbol{Z}} \in \mathcal{H}. \tag{16}$$

We also need to define the expectation of $g_{\boldsymbol{Z}}$ as

$$g = \mathbb{E} g_{\boldsymbol{Z}} = \int_{\mathcal{X}} k(\boldsymbol{x}, \cdot) f_\rho^*(\boldsymbol{x}) d\mu(\boldsymbol{x}) = S_k^* f_\rho^* \in \mathcal{H},$$

and

$$f_\lambda = \left(T + \lambda\right)^{-1} g = \left(T + \lambda\right)^{-1} S_k^* f_\rho^*. \tag{17}$$

Denoting $\epsilon_i = y_i - f_\rho^*(\boldsymbol{x}_i)$, we have the decomposition

$$\begin{aligned}
\hat{f}_\lambda - f_\rho^* &= \frac{1}{n}(T_{\boldsymbol{X}} + \lambda)^{-1} \sum_{i=1}^n K_{\boldsymbol{x}_i} y_i - f_\rho^* \\
&= \frac{1}{n}(T_{\boldsymbol{X}} + \lambda)^{-1} \sum_{i=1}^n K_{\boldsymbol{x}_i} (f_\rho^*(\boldsymbol{x}_i) + \epsilon_i) - f_\rho^* \\
&= (T_{\boldsymbol{X}} + \lambda)^{-1} \tilde{g}_{\boldsymbol{Z}} + \frac{1}{n} \sum_{i=1}^n (T_{\boldsymbol{X}} + \lambda)^{-1} K_{\boldsymbol{x}_i} \epsilon_i - f_\rho^* \\
&= \left(\tilde{f}_\lambda - f_\rho^*\right) + \frac{1}{n} \sum_{i=1}^n (T_{\boldsymbol{X}} + \lambda)^{-1} K_{\boldsymbol{x}_i} \epsilon_i.
\end{aligned}$$

Taking expectation over the noise ϵ_i conditioned on $\boldsymbol{X}$ and noticing that $\epsilon_i | \boldsymbol{x}$ are independent noises with mean 0 and variance σ^2, we obtain the bias-variance decomposition:

$$\mathbb{E}\left(\left\| \hat{f}_\lambda - f_\rho^* \right\|_{L^2}^2 \;\middle|\; \boldsymbol{X} \right) = \mathbf{Bias}^2(\lambda) + \mathbf{Var}(\lambda), \tag{18}$$

where

$$\mathbf{Bias}^2(\lambda) := \left\| \tilde{f}_\lambda - f_\rho^* \right\|_{L^2}^2, \quad \mathbf{Var}(\lambda) := \frac{\sigma^2}{n^2} \sum_{i=1}^n \left\| (T_{\boldsymbol{X}} + \lambda)^{-1} k(\boldsymbol{x}_i, \cdot) \right\|_{L^2}^2. \tag{19}$$

Given the decomposition (18), we next derive the upper and lower bounds of $\mathbf{Bias}^2(\lambda)$ and $\mathbf{Var}(\lambda)$ in the following two subsections.

A.2 Variance term

In this subsection, our goal is to derive Theorem 15, which shows the upper and lower bounds of variance under some approximation conditions. Before formally introducing Theorem 15, we need a lot of preparatory work.

Following Li et al. (2024), we consider the sample subspace

$$\mathcal{H}_n = \operatorname{span}\{k(\boldsymbol{x}_1, \cdot), \ldots, k(\boldsymbol{x}_n, \cdot)\} \subset \mathcal{H}.$$

Recall the notation $\mathbb{K}(\boldsymbol{X}, \boldsymbol{X}) = \left(k\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)\right)_{n \times n}$ and $\mathbb{K}(\boldsymbol{X}, \cdot) = \{k(\boldsymbol{x}_1, \cdot), \ldots, k(\boldsymbol{x}_n, \cdot)\}$. Define the normalized sample kernel matrix

$$\boldsymbol{K} = \frac{1}{n} \mathbb{K}(\boldsymbol{X}, \boldsymbol{X}).$$

Then, it is easy to verify that $\mathrm{Ran}(T_{\boldsymbol{X}}) = \mathcal{H}_n$ and $\boldsymbol{K}$ is the representation matrix of $T_{\boldsymbol{X}}$ under the natural basis $\{k(\boldsymbol{x}_1, \cdot), \ldots, k(\boldsymbol{x}_n, \cdot)\}$. Consequently, for any continuous function φ we have

$$\varphi(T_{\boldsymbol{X}})\mathbb{K}(\boldsymbol{X}, \cdot) = \varphi(\boldsymbol{K})\mathbb{K}(X, \cdot), \tag{20}$$

where the left-hand side means applying the operator elementwise. Since the property of RKHS implies $\langle k(\boldsymbol{x}, \cdot), f\rangle_{\mathcal{H}} = f(\boldsymbol{x}),\ \forall f \in \mathcal{H}$, taking inner product elementwise between (20) and f, we have

$$(\varphi(T_{\boldsymbol{X}})f)[\boldsymbol{X}] = \varphi(\boldsymbol{K})f[\boldsymbol{X}]. \tag{21}$$

Moreover, for $f, g \in \mathcal{H}$, we define empirical semi-inner product

$$\langle f, g\rangle_{L^2,n} = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i)g(\boldsymbol{x}_i) = \frac{1}{n} f[\boldsymbol{X}]^\top g[\boldsymbol{X}], \tag{22}$$

and denote by $\|\cdot\|_{L^2,n}$ the corresponding empirical semi-norm. We also denote by P_n the empirical measure with respect to $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$.

For simplicity of notation, we denote $k_{\boldsymbol{x}}(\cdot) = k(\boldsymbol{x}, \cdot)$, $\boldsymbol{x} \in \mathcal{X}$ in the rest of the proof. The following lemma rewrites the variance term (19) using the empirical semi-norm.

Lemma 11 *The variance term in* (19) *satisfies that*

$$\mathbf{Var}(\lambda) = \frac{\sigma^2}{n}\int_{\mathcal{X}} \left\|(T_{\boldsymbol{X}} + \lambda)^{-1} k_{\boldsymbol{x}}(\cdot)\right\|_{L^2,n}^2 \mathrm{d}\mu(\boldsymbol{x}). \tag{23}$$

Proof First, we have

$$\begin{aligned}
\mathbf{Var}(\lambda) &\coloneqq \frac{\sigma^2}{n^2}\sum_{i=1}^{n} \left\|(T_{\boldsymbol{X}} + \lambda)^{-1} k(\boldsymbol{x}_i, \cdot)\right\|_{L^2}^2 \\
&= \frac{\sigma^2}{n^2} \left\|(T_{\boldsymbol{X}} + \lambda)^{-1}\mathbb{K}(\boldsymbol{X}, \cdot)\right\|_{L^2(\mathbb{R}^n)}^2 \\
&\overset{(20)}{=} \frac{\sigma^2}{n^2} \left\|(\boldsymbol{K} + \lambda)^{-1}\mathbb{K}(\boldsymbol{X}, \cdot)\right\|_{L^2(\mathbb{R}^n)}^2 \\
&= \frac{\sigma^2}{n^2}\int_{\mathcal{X}} \mathbb{K}(\boldsymbol{x}, \boldsymbol{X})(\boldsymbol{K} + \lambda)^{-2}\mathbb{K}(\boldsymbol{X}, \boldsymbol{x})\ \mathrm{d}\mu(\boldsymbol{x}).
\end{aligned} \tag{24}$$

Next, using (21) and the fact that $k_{\boldsymbol{x}}[\boldsymbol{X}] = \mathbb{K}(\boldsymbol{X}, \boldsymbol{x})$, we have

$$\left((T_{\boldsymbol{X}} + \lambda)^{-1} k_{\boldsymbol{x}}\right)[\boldsymbol{X}] = (\boldsymbol{K} + \lambda)^{-1} k_{\boldsymbol{x}}[\boldsymbol{X}] = (\boldsymbol{K} + \lambda)^{-1}\mathbb{K}(\boldsymbol{X}, \boldsymbol{x}),$$

which implies

$$\begin{aligned}
\frac{1}{n}\mathbb{K}(\boldsymbol{x}, \boldsymbol{X})(\boldsymbol{K} + \lambda)^{-2}\mathbb{K}(\boldsymbol{X}, \boldsymbol{x}) &= \frac{1}{n}\left\|(\boldsymbol{K} + \lambda)^{-1}\mathbb{K}(\boldsymbol{X}, \boldsymbol{x})\right\|_{\mathbb{R}^n}^2 \\
&= \frac{1}{n}\left\|\left((T_{\boldsymbol{X}} + \lambda)^{-1} k_{\boldsymbol{x}}\right)[\boldsymbol{X}]\right\|_{\mathbb{R}^n}^2 \\
&= \left\|(T_{\boldsymbol{X}} + \lambda)^{-1} k_{\boldsymbol{x}}\right\|_{L^2,n}^2.
\end{aligned} \tag{25}$$

Therefore, plugging (25) into (24), we obtain the desired results

$$\mathbf{Var}(\lambda) = \frac{\sigma^2}{n} \int_{\mathcal{X}} \left\| (T_{\boldsymbol{X}} + \lambda)^{-1} k_{\boldsymbol{x}}(\cdot) \right\|_{L^2,n}^2 \mathrm{d}\mu(\boldsymbol{x}).$$

■

The operator form (23) allows us to apply concentration inequalities and establish the following two-step approximation (recall the notations $T_{\boldsymbol{X}\lambda}$ and T_λ in (14)).

$$\int_{\mathcal{X}} \left\| T_{\boldsymbol{X}\lambda}^{-1} k_{\boldsymbol{x}} \right\|_{L^2,n}^2 \mathrm{d}\mu(\boldsymbol{x}) \overset{A}{\approx} \int_{\mathcal{X}} \left\| T_\lambda^{-1} k_{\boldsymbol{x}} \right\|_{L^2,n}^2 \mathrm{d}\mu(\boldsymbol{x}) \overset{B}{\approx} \int_{\mathcal{X}} \left\| T_\lambda^{-1} k_{\boldsymbol{x}} \right\|_{L^2}^2 \mathrm{d}\mu(\boldsymbol{x}).$$

Note that the above two-step approximation is an enhanced version of approximation (S24) in Li et al. (2024).

Approximation B. The following lemma characterizes the magnitude of Approximation B in high probability. Recall the definitions of $\mathcal{N}_1(\lambda)$ and $\mathcal{N}_2(\lambda)$ in (4).

Lemma 12 (Approximation B) *Suppose that Assumption 1, 2 and 3 hold. For any $\lambda = \lambda(d,n) \to 0$ and any fixed $\delta \in (0,1)$, when n is sufficiently large, with probability at least $1-\delta$, we have*

$$\frac{1}{2} \int_{\mathcal{X}} \left\| T_\lambda^{-1} k_{\boldsymbol{x}} \right\|_{L^2}^2 \mathrm{d}\mu(\boldsymbol{x}) - R_2 \le \int_{\mathcal{X}} \left\| T_\lambda^{-1} k_{\boldsymbol{x}} \right\|_{L^2,n}^2 \mathrm{d}\mu(\boldsymbol{x}) \le \frac{3}{2} \int_{\mathcal{X}} \left\| T_\lambda^{-1} k_{\boldsymbol{x}} \right\|_{L^2}^2 \mathrm{d}\mu(\boldsymbol{x}) + R_2, \tag{26}$$

where

$$R_2 = \frac{5\mathcal{N}_2(\lambda)}{3n} \ln \frac{2}{\delta}.$$

Proof Define a function

$$\begin{aligned} f(\boldsymbol{z}) &= \int_{\mathcal{X}} \left(T_\lambda^{-1} k_{\boldsymbol{x}}(\boldsymbol{z}) \right)^2 \mathrm{d}\mu(\boldsymbol{x}) \\ &= \int_{\mathcal{X}} \sum_{i=1}^{\infty} \left(\frac{\lambda_i}{\lambda_i + \lambda} \right)^2 e_i^2(\boldsymbol{x}) e_i^2(\boldsymbol{z}) \mathrm{d}\mu(\boldsymbol{x}) \\ &= \sum_{i=1}^{\infty} \left(\frac{\lambda_i}{\lambda_i + \lambda} \right)^2 e_i^2(\boldsymbol{z}). \end{aligned}$$

Since Assumption 3 holds, we have

$$\|f\|_{L^\infty} \le \sum_{i=1}^{\infty} \left(\frac{\lambda_i}{\lambda_i + \lambda} \right)^2 = \mathcal{N}_2(\lambda); \quad \|f\|_{L^1} = \sum_{i=1}^{\infty} \left(\frac{\lambda_i}{\lambda_i + \lambda} \right)^2 = \mathcal{N}_2(\lambda).$$

Applying Proposition 36 for $\sqrt{f}$ and noticing that $\|\sqrt{f}\|_{L^\infty} = \sqrt{\|f\|_{L^\infty}} = \mathcal{N}_2(\lambda)^{\frac{1}{2}}$, we have

$$\frac{1}{2} \left\| \sqrt{f} \right\|_{L^2}^2 - \frac{5\mathcal{N}_2(\lambda)}{3n} \ln \frac{2}{\delta} \le \left\| \sqrt{f} \right\|_{L^2,n}^2 \le \frac{3}{2} \left\| \sqrt{f} \right\|_{L^2}^2 + \frac{5\mathcal{N}_2(\lambda)}{3n} \ln \frac{2}{\delta}, \tag{27}$$

with probability at least $1-\delta$.

On the one hand, we have

$$\begin{aligned}\left\|\sqrt{f}\right\|_{L^2,n}^2 &= \int_{\mathcal{X}} f(\boldsymbol{y})\mathrm{d}P_n(\boldsymbol{y}) = \int_{\mathcal{X}}\left[\int_{\mathcal{X}}\left(T_\lambda^{-1}k_{\boldsymbol{x}}(\boldsymbol{y})\right)^2\mathrm{d}\mu(\boldsymbol{x})\right]\mathrm{d}P_n(\boldsymbol{y})\\ &= \int_{\mathcal{X}}\left[\int_{\mathcal{X}}\left(T_\lambda^{-1}k_{\boldsymbol{x}}(\boldsymbol{y})\right)^2\mathrm{d}P_n(\boldsymbol{y})\right]\mathrm{d}\mu(\boldsymbol{x})\\ &= \int_{\mathcal{X}}\left\|T_\lambda^{-1}k_{\boldsymbol{x}}\right\|_{L^2,n}^2\mathrm{d}\mu(\boldsymbol{x}).\end{aligned}$$

On the other hand, we have

$$\begin{aligned}\left\|\sqrt{f}\right\|_{L^2}^2 &= \int_{\mathcal{X}} f(\boldsymbol{z})\mathrm{d}\mu(\boldsymbol{z})\\ &= \int_{\mathcal{X}}\left[\int_{\mathcal{X}}\left(T_\lambda^{-1}k_{\boldsymbol{x}}(\boldsymbol{z})\right)^2\mathrm{d}\mu(\boldsymbol{x})\right]\mathrm{d}\mu(\boldsymbol{z})\\ &= \int_{\mathcal{X}}\left\|T_\lambda^{-1}k_{\boldsymbol{x}}\right\|_{L^2}^2\mathrm{d}\mu(\boldsymbol{x}).\end{aligned}$$

Therefore, the proof follows from (27). ■

Approximation A. The proof of Approximation A requires the following proposition, which is a simple but important observation by Li et al. (2023a).

Proposition 13 *For any $f, g \in \mathcal{H}$, we have*

$$\langle f, g\rangle_{L^2,n} = \langle T_{\boldsymbol{X}} f, g\rangle_{\mathcal{H}} = \langle T_{\boldsymbol{X}}^{1/2} f, T_{\boldsymbol{X}}^{1/2} g\rangle_{\mathcal{H}}.$$

Proof By the definition of $T_{\boldsymbol{X}}$, we have $T_{\boldsymbol{X}} f = \frac{1}{n}\sum_{i=1}^n f(\boldsymbol{x}_i)k(\boldsymbol{x}_i,\cdot)$, and thus

$$\langle T_{\boldsymbol{X}} f, g\rangle_{\mathcal{H}} = \frac{1}{n}\sum_{i=1}^n f(\boldsymbol{x}_i)\langle k(\boldsymbol{x}_i,\cdot), g\rangle_{\mathcal{H}} = \frac{1}{n}\sum_{i=1}^n f(\boldsymbol{x}_i)g(\boldsymbol{x}_i) = \langle f, g\rangle_{L^2,n}.$$

The second inequality comes from the definition of $T_{\boldsymbol{X}}^{1/2}$. ■

The following lemma characterizes the magnitude of Approximation A.

Lemma 14 (Approximation A) *Suppose that Assumption 1, 2 and 3 hold. Define $R_1 = R_1(\lambda, \boldsymbol{X})$ as a function of λ and $\boldsymbol{X}$:*

$$R_1 \coloneqq \left|\int_{\mathcal{X}}\left\|T_{\boldsymbol{X}\lambda}^{-1}k_{\boldsymbol{x}}\right\|_{L^2,n}^2\mathrm{d}\mu(\boldsymbol{x}) - \int_{\mathcal{X}}\left\|T_\lambda^{-1}k_{\boldsymbol{x}}\right\|_{L^2,n}^2\mathrm{d}\mu(\boldsymbol{x})\right|. \tag{28}$$

Suppose that $\lambda = \lambda(d,n)$ satisfies $\frac{\mathcal{N}_1(\lambda)}{n}\ln n = o(1)$. Then for any fixed $\delta \in (0,1)$, when n is sufficiently large, with probability at least $1-\delta$, we have

$$R_1 \le 36n^{-1}\mathcal{N}_1(\lambda)^2\ln n + 12n^{-\frac{1}{2}}\mathcal{N}_1(\lambda)\mathcal{N}_2(\lambda)^{\frac{1}{2}}(\ln n)^{\frac{1}{2}}.$$

Proof First, we rewrite R_1 as

$$\begin{aligned}
R_1 &= \left| \int_{\mathcal{X}} \left\| T_{\boldsymbol{X}\lambda}^{-1} k_{\boldsymbol{x}} \right\|_{L^2,n}^2 \mathrm{d}\mu(\boldsymbol{x}) - \int_{\mathcal{X}} \left\| T_{\lambda}^{-1} k_{\boldsymbol{x}} \right\|_{L^2,n}^2 \mathrm{d}\mu(\boldsymbol{x}) \right| \\
&\leq \int_{\mathcal{X}} \left| \left\| T_{\boldsymbol{X}}^{\frac{1}{2}} T_{\boldsymbol{X}\lambda}^{-1} k_{\boldsymbol{x}} \right\|_{\mathcal{H}}^2 - \left\| T_{\boldsymbol{X}}^{\frac{1}{2}} T_{\lambda}^{-1} k_{\boldsymbol{x}} \right\|_{\mathcal{H}}^2 \right| \mathrm{d}\mu(\boldsymbol{x}) \\
&= \int_{\mathcal{X}} \left| \left\| T_{\boldsymbol{X}}^{\frac{1}{2}} T_{\boldsymbol{X}\lambda}^{-1} k_{\boldsymbol{x}} \right\|_{\mathcal{H}} - \left\| T_{\boldsymbol{X}}^{\frac{1}{2}} T_{\lambda}^{-1} k_{\boldsymbol{x}} \right\|_{\mathcal{H}} \right| \cdot \left| \left\| T_{\boldsymbol{X}}^{\frac{1}{2}} T_{\boldsymbol{X}\lambda}^{-1} k_{\boldsymbol{x}} \right\|_{\mathcal{H}} + \left\| T_{\boldsymbol{X}}^{\frac{1}{2}} T_{\lambda}^{-1} k_{\boldsymbol{x}} \right\|_{\mathcal{H}} \right| \mathrm{d}\mu(\boldsymbol{x}). \\
&:= \int_{\mathcal{X}} |X_1 - X_2| \cdot |X_1 + X_2| \, \mathrm{d}\mu(\boldsymbol{x}). \qquad (29)
\end{aligned}$$

where for the second line, we use Proposition 13 to transfer $\|\cdot\|_{L^2,n}$ norms into $\|\cdot\|_{\mathcal{H}}$.

(I) : Given the notations of X_1, X_2 in (29) (both are functions of $\boldsymbol{x}, \boldsymbol{X}$ and λ), we begin to handle $|X_1 - X_2|$:

$$\begin{aligned}
|X_1 - X_2| &\leq \left\| T_{\boldsymbol{X}}^{1/2} T_{\boldsymbol{X}\lambda}^{-1} \left(T - T_{\boldsymbol{X}} \right) T_{\lambda}^{-1} k_{\boldsymbol{x}} \right\|_{\mathcal{H}} \\
&\leq \left\| T_{\boldsymbol{X}}^{1/2} T_{\boldsymbol{X}\lambda}^{-1/2} \right\| \cdot \left\| T_{\boldsymbol{X}\lambda}^{-1/2} T_{\lambda}^{1/2} \right\| \cdot \left\| T_{\lambda}^{-1/2} \left(T - T_{\boldsymbol{X}} \right) T_{\lambda}^{-1/2} \right\| \cdot \left\| T_{\lambda}^{-1/2} k_{\boldsymbol{x}} \right\|_{\mathcal{H}} \qquad (30)
\end{aligned}$$

(i) Now we bound the third term in (30). Denote $A_i = T_{\lambda}^{-\frac{1}{2}} (T - T_{\boldsymbol{x}_i}) T_{\lambda}^{-\frac{1}{2}}$, using Lemma 40, we have

$$\|A_i\| \leq \|T_{\lambda}^{-\frac{1}{2}} T T_{\lambda}^{-\frac{1}{2}}\| + \|T_{\lambda}^{-\frac{1}{2}} T_{\boldsymbol{x}_i} T_{\lambda}^{-\frac{1}{2}}\| \leq 2\mathcal{N}_1(\lambda), \quad \mu\text{-a.e. } \boldsymbol{x} \in \mathcal{X}.$$

We use $A \preceq B$ to denote that $A - B$ is a positive semi-definite operator. Using the fact that $\mathbb{E}(B - \mathbb{E}B)^2 \preceq \mathbb{E}B^2$ for a self-adjoint operator B, we have

$$\mathbb{E}A_i^2 \preceq \mathbb{E}\left[T_{\lambda}^{-\frac{1}{2}} T_{\boldsymbol{x}_i} T_{\lambda}^{-\frac{1}{2}} \right]^2.$$

In addition, Lemma 40 shows that $0 \preceq T_{\lambda}^{-\frac{1}{2}} T_{\boldsymbol{x}_i} T_{\lambda}^{-\frac{1}{2}} \preceq \mathcal{N}_1(\lambda), \mu$-a.e. $\boldsymbol{x} \in \mathcal{X}$. So we have

$$\mathbb{E}A_i^2 \preceq \mathbb{E}\left[T_{\lambda}^{-\frac{1}{2}} T_{\boldsymbol{x}_i} T_{\lambda}^{-\frac{1}{2}} \right]^2 \preceq \mathbb{E}\left[\mathcal{N}_1(\lambda) \cdot T_{\lambda}^{-\frac{1}{2}} T_{\boldsymbol{x}_i} T_{\lambda}^{-\frac{1}{2}} \right] = \mathcal{N}_1(\lambda) T_{\lambda}^{-1} T,$$

Defining an operator $V := \mathcal{N}_1(\lambda) T_{\lambda}^{-1} T$, we have

$$\begin{aligned}
\|V\| &= \mathcal{N}_1(\lambda) \frac{\lambda_1}{\lambda_1 + \lambda} = \mathcal{N}_1(\lambda) \frac{\|T\|}{\|T\| + \lambda} \leq \mathcal{N}_1(\lambda); \\
\mathrm{tr}V &= \mathcal{N}_1(\lambda)^2; \\
\frac{\mathrm{tr}V}{\|V\|} &= \frac{\mathcal{N}_1(\lambda)(\|T\| + \lambda)}{\|T\|}.
\end{aligned}$$

Applying Lemma 37 to A_i, V, for any fixed $\delta \in (0,1)$, with probability at least $1-\delta$, we have

$$\|T_\lambda^{-\frac{1}{2}}(T - T_{\boldsymbol{X}})T_\lambda^{-\frac{1}{2}}\| \le \frac{4\mathcal{N}_1(\lambda)}{3n}\beta + \sqrt{\frac{2\mathcal{N}_1(\lambda)}{n}\beta},$$

where

$$\beta = \ln \frac{4\mathcal{N}_1(\lambda)(\|T\| + \lambda)}{\delta\|T\|}.$$

Further recall that the condition $\mathcal{N}_1(\lambda)\ln n/n = o(1)$ implies that $\mathcal{N}_1(\lambda) = O(n)$ and thus $\beta = O(\ln n)$, so when n is sufficiently large, we conclude that

$$\|T_\lambda^{-\frac{1}{2}}(T - T_{\boldsymbol{X}})T_\lambda^{-\frac{1}{2}}\| \le \sqrt{\frac{2\mathcal{N}_1(\lambda)}{n}\beta} \le n^{-\frac{1}{2}}\mathcal{N}_1(\lambda)^{\frac{1}{2}}(\ln n)^{\frac{1}{2}}. \tag{31}$$

(ii) Next we bound the forth term in (30). Using Lemma 39, we have

$$\|T_\lambda^{-\frac{1}{2}}k_{\boldsymbol{x}}\|_{\mathcal{H}} \le \mathcal{N}_1(\lambda)^{\frac{1}{2}}, \quad \mu\text{-a.e. } \boldsymbol{x} \in \mathcal{X}. \tag{32}$$

(iii) Finally, we bound the first two terms in (30). Since we have assumed $\mathcal{N}_1(\lambda)\ln n/n = o(1)$, (31) implies that when n is sufficiently large,

$$a := \|T_\lambda^{-\frac{1}{2}}(T - T_{\boldsymbol{X}})T_\lambda^{-\frac{1}{2}}\| \le \frac{2}{3}.$$

Therefore, we have

$$\begin{aligned}
\left\|T_\lambda^{-\frac{1}{2}}T_{\boldsymbol{X}\lambda}^{\frac{1}{2}}\right\|^2 &= \left\|T_\lambda^{-\frac{1}{2}}T_{\boldsymbol{X}\lambda}T_\lambda^{-\frac{1}{2}}\right\| = \left\|T_\lambda^{-\frac{1}{2}}\left(T_{\boldsymbol{X}} + \lambda\right)T_\lambda^{-\frac{1}{2}}\right\| \\
&= \left\|T_\lambda^{-\frac{1}{2}}\left(T_{\boldsymbol{X}} - T + T + \lambda\right)T_\lambda^{-\frac{1}{2}}\right\| \\
&= \left\|T_\lambda^{-\frac{1}{2}}\left(T_{\boldsymbol{X}} - T\right)T_\lambda^{-\frac{1}{2}} + I\right\| \\
&\le a + 1 \le 2;
\end{aligned} \tag{33}$$

and

$$\begin{aligned}
\left\|T_\lambda^{\frac{1}{2}}T_{\boldsymbol{X}\lambda}^{-\frac{1}{2}}\right\|^2 &= \left\|T_\lambda^{\frac{1}{2}}T_{\boldsymbol{X}\lambda}^{-1}T_\lambda^{\frac{1}{2}}\right\| = \left\|\left(T_\lambda^{-\frac{1}{2}}T_{\boldsymbol{X}\lambda}T_\lambda^{-\frac{1}{2}}\right)^{-1}\right\| \\
&= \left\|\left(I - T_\lambda^{-\frac{1}{2}}(T_{\boldsymbol{X}} - T)T_\lambda^{-\frac{1}{2}}\right)^{-1}\right\| \\
&\le \sum_{k=0}^{\infty}\left\|T_\lambda^{-\frac{1}{2}}(T_{\boldsymbol{X}} - T)T_\lambda^{-\frac{1}{2}}\right\|^k \\
&\le \sum_{k=0}^{\infty}\left(\frac{2}{3}\right)^k \le 3.
\end{aligned} \tag{34}$$

Plugging (31), (32), (33) and (34), into (30), when n is sufficiently large, with probability at least $1-\delta$, we have

$$|X_1 - X_2| \leq 6n^{-\frac{1}{2}}\mathcal{N}_1(\lambda)\left(\ln n\right)^{\frac{1}{2}}, \quad \mu\text{-a.e. } \boldsymbol{x} \in \mathcal{X}. \tag{35}$$

(II) : Furthermore, when n is sufficiently large, with probability at least $1-\delta$, we also have

$$\begin{aligned}
\int_{\mathcal{X}} X_2 \mathrm{d}\mu(\boldsymbol{x}) &= \int_{\mathcal{X}} \left\|T_\lambda^{-1} k_{\boldsymbol{x}}\right\|_{L^2,n} \mathrm{d}\mu(\boldsymbol{x}) \\
&\leq \left[\int_{\mathcal{X}} \left\|T_\lambda^{-1} k_{\boldsymbol{x}}\right\|_{L^2,n}^2 \mathrm{d}\mu(\boldsymbol{x})\right]^{\frac{1}{2}} \\
&\leq \left(\frac{3}{2}\mathcal{N}_2(\lambda) + R_2\right)^{\frac{1}{2}} \\
&\leq \left(2\mathcal{N}_2(\lambda)\right)^{\frac{1}{2}},
\end{aligned} \tag{36}$$

where the third line follows from Lemma 12.

Now we are ready to derive the upper bound of R_1 in (29). Combining the bound (35) and (36), when n is sufficiently large, with probability at least $1-2\delta$, we have

$$\begin{aligned}
R_1 &\leq \int_{\mathcal{X}} |X_1 - X_2| \cdot |X_1 + X_2| \, \mathrm{d}\mu(\boldsymbol{x}) \\
&= \int_{\mathcal{X}} |X_1 - X_2| \cdot |X_1 - X_2 + 2X_2| \, \mathrm{d}\mu(\boldsymbol{x}) \\
&\leq \int_{\mathcal{X}} |X_1 - X_2|^2 \, \mathrm{d}\mu(\boldsymbol{x}) + \int_{\mathcal{X}} |X_1 - X_2| \cdot 2X_2 \, \mathrm{d}\mu(\boldsymbol{x}) \\
&\leq 36n^{-1}\mathcal{N}_1(\lambda)^2 \ln n + 24n^{-\frac{1}{2}}\mathcal{N}_1(\lambda)\mathcal{N}_2(\lambda)^{\frac{1}{2}} \left(\ln n\right)^{\frac{1}{2}}.
\end{aligned} \tag{37}$$

Without loss of generality, we can assume (37) holds with probability at least $1-\delta$ and we finish the proof. ■

Final proof of the variance term. Now we are ready to state the theorem about the variance term.

Theorem 15 (Variance term) *Suppose that Assumption 1, 2 and 3 hold. If the following approximation conditions hold for some $\lambda = \lambda(d,n) \to 0$:*

$$\frac{\mathcal{N}_1(\lambda)}{n} \ln n = o(1); \quad n^{-1}\mathcal{N}_1(\lambda)^2 \ln n = o(\mathcal{N}_2(\lambda)), \tag{38}$$

then we have

$$\mathbf{Var}(\lambda) = \Theta_{\mathbb{P}}\left(\frac{\sigma^2 \mathcal{N}_2(\lambda)}{n}\right).$$

Proof Lemma 11 has shown that

$$\mathbf{Var}(\lambda) = \frac{\sigma^2}{n} \int_{\mathcal{X}} \left\|(T_{\boldsymbol{X}} + \lambda)^{-1} k_{\boldsymbol{x}}(\cdot)\right\|_{L^2,n}^2 \mathrm{d}\mu(\boldsymbol{x}).$$

Denote R_1 as in Lemma 14, then conditions (38) and Lemma 14 imply that

$$R_1 = o_{\mathbb{P}}\left(\mathcal{N}_2(\lambda)\right).$$

Further recalling that in Lemma 12, we have defined

$$R_2 = \frac{5\mathcal{N}_2(\lambda)}{3n}\ln\frac{2}{\delta}.$$

On the one hand, Lemma 12 shows that, for any $\delta \in (0,1)$, when n is sufficiently large, with probability at least $1-\delta$, we have

$$\begin{aligned} n\mathbf{Var}(\lambda)/\sigma^2 = \int_{\mathcal{X}} \left\|T_{\boldsymbol{X}\lambda}^{-1} k_{\boldsymbol{x}}\right\|_{L^2,n}^2 \mathrm{d}\mu(\boldsymbol{x}) &\le \int_{\mathcal{X}} \left\|T_{\lambda}^{-1} k_{\boldsymbol{x}}\right\|_{L^2,n}^2 \mathrm{d}\mu(\boldsymbol{x}) + R_1 \\ &\le \frac{3}{2}\int_{\mathcal{X}} \left\|T_{\lambda}^{-1} k_{\boldsymbol{x}}\right\|_{L^2}^2 \mathrm{d}\mu(\boldsymbol{x}) + R_1 + R_2 \\ &= \frac{3}{2}\mathcal{N}_2(\lambda) + R_1 + R_2, \end{aligned}$$

which further implies

$$n\mathbf{Var}(\lambda)/\sigma^2 = O_{\mathbb{P}}\left(\mathcal{N}_2(\lambda)\right). \tag{39}$$

On the other hand, we also have

$$\begin{aligned} n\mathbf{Var}(\lambda)/\sigma^2 = \int_{\mathcal{X}} \left\|T_{\boldsymbol{X}\lambda}^{-1} k_{\boldsymbol{x}}\right\|_{L^2,n}^2 \mathrm{d}\mu(\boldsymbol{x}) &\ge \int_{\mathcal{X}} \left\|T_{\lambda}^{-1} k_{\boldsymbol{x}}\right\|_{L^2,n}^2 \mathrm{d}\mu(\boldsymbol{x}) - R_1 \\ &\ge \frac{1}{2}\int_{\mathcal{X}} \left\|T_{\lambda}^{-1} k_{\boldsymbol{x}}\right\|_{L^2}^2 \mathrm{d}\mu(\boldsymbol{x}) - R_1 - R_2 \\ &= \frac{1}{2}\mathcal{N}_2(\lambda) - R_1 - R_2, \end{aligned}$$

which further implies

$$n\mathbf{Var}(\lambda)/\sigma^2 = \Omega_{\mathbb{P}}\left(\mathcal{N}_2(\lambda)\right). \tag{40}$$

Combining (39) and (40), we finish the proof. ■

A.3 Bias term

In this subsection, our goal is to derive Theorem 18, which shows the upper and lower bounds of bias under some approximation conditions.

The triangle inequality implies that

$$\mathbf{Bias}(\lambda) = \left\|\tilde{f}_\lambda - f_\rho^*\right\|_{L^2} \ge \left\|f_\lambda - f_\rho^*\right\|_{L^2} - \left\|\tilde{f}_\lambda - f_\lambda\right\|_{L^2}, \tag{41}$$

where $\tilde{f}_\lambda, f_\lambda$ are defined as (16) and (17).

The following lemma characterizes the dominant term of $\mathbf{Bias}(\lambda)$.

Lemma 16 *Suppose that Assumption 1 and 2 hold. Then for any $\lambda = \lambda(d,n) \to 0$,*

$$\left\|f_\lambda - f_\rho^*\right\|_{L^2} = \mathcal{M}_2(\lambda)^{\frac{1}{2}}. \tag{42}$$

Proof Recall that we have defined $f_\rho^* = \sum_{i=1}^{\infty} f_i e_i(\boldsymbol{x}) \in L^2(\mathcal{X},\mu)$ and $f_\lambda = (T+\lambda)^{-1} S_k^* f_\rho^*$. Therefore, we have

$$\begin{aligned}
\left\|f_\lambda - f_\rho^*\right\|_{L^2}^2 &= \left\|\sum_{i=1}^{\infty} f_i e_i(\boldsymbol{x}) - \sum_{i=1}^{\infty} \frac{\lambda_i}{\lambda_i+\lambda} f_i e_i(\boldsymbol{x})\right\|_{L^2}^2 \\
&= \left\|\sum_{i=1}^{\infty} \frac{\lambda}{\lambda_i+\lambda} f_i e_i(\boldsymbol{x})\right\|_{L^2}^2 \\
&= \sum_{i=1}^{\infty} \left(\frac{\lambda}{\lambda_i+\lambda} f_i\right)^2 \\
&= \mathcal{M}_2(\lambda).
\end{aligned}$$

■

Our next goal is to prove that second term in (41) is higher order infinitesimal, i.e., $\left\|\tilde{f}_\lambda - f_\lambda\right\|_{L^2} = o_{\mathbb{P}}(\mathcal{M}_2(\lambda)^{\frac{1}{2}})$.

Lemma 17 *Suppose that Assumption 1, 2 and 3 hold. If the following approximation conditions hold for some $\lambda = \lambda(d,n) \to 0$:*

$$\frac{\mathcal{N}_1(\lambda)}{n} \ln n = o(1); \quad n^{-1}\mathcal{N}_1(\lambda)^{\frac{1}{2}}\mathcal{M}_1(\lambda) = o\left(\mathcal{M}_2(\lambda)^{\frac{1}{2}}\right), \tag{43}$$

then we have

$$\left\|\tilde{f}_\lambda - f_\lambda\right\|_{L^2} = o_{\mathbb{P}}(\mathcal{M}_2(\lambda)^{\frac{1}{2}}).$$

Proof To begin with, by definition, we rewrite $\left\|\tilde{f}_\lambda - f_\lambda\right\|_{L_2}$ as follows

$$\begin{aligned}
\left\|\tilde{f}_\lambda - f_\lambda\right\|_{L_2} &= \left\|S_k\left(\tilde{f}_\lambda - f_\lambda\right)\right\|_{L^2} \\
&= \left\|S_k T_\lambda^{-\frac{1}{2}} \cdot T_\lambda^{\frac{1}{2}} T_{\boldsymbol{X}\lambda}^{-1} T_\lambda^{\frac{1}{2}} \cdot T_\lambda^{-\frac{1}{2}} T_{\boldsymbol{X}\lambda}\left(\tilde{f}_\lambda - f_\lambda\right)\right\|_{L^2} \\
&\le \left\|S_k T_\lambda^{-\frac{1}{2}}\right\|_{\mathscr{B}(\mathcal{H},L^2)} \cdot \left\|T_\lambda^{\frac{1}{2}} T_{\boldsymbol{X}\lambda}^{-1} T_\lambda^{\frac{1}{2}}\right\|_{\mathscr{B}(\mathcal{H},\mathcal{H})} \cdot \left\|T_\lambda^{-\frac{1}{2}}\left(\tilde{g}_{\boldsymbol{Z}} - T_{\boldsymbol{X}\lambda} f_\lambda\right)\right\|_{\mathcal{H}}.
\end{aligned} \tag{44}$$

(i) For any $f \in \mathcal{H}$ and $\|f\|_{\mathcal{H}} = 1$, suppose that $f = \sum_{i=1}^{\infty} a_i \lambda_i^{1/2} e_i$ satisfying that $\sum_{i=1}^{\infty} a_i^2 = 1$. So for the first term in (44), we have

$$\begin{aligned}
\left\| S_k T_\lambda^{-\frac{1}{2}} \right\|_{\mathscr{B}(\mathcal{H}, L^2)} &= \sup_{\|f\|_{\mathcal{H}}=1} \left\| S_k T_\lambda^{-\frac{1}{2}} f \right\|_{L^2} \\
&\le \sup_{\|f\|_{\mathcal{H}}=1} \left\| \sum_{i=1}^{\infty} \frac{\lambda_i^{\frac{1}{2}}}{(\lambda_i + \lambda)^{\frac{1}{2}}} a_i e_i \right\|_{L^2} \\
&\le \sup_{i \ge 1} \frac{\lambda_i^{\frac{1}{2}}}{(\lambda_i + \lambda)^{\frac{1}{2}}} \cdot \sup_{\|f\|_{\mathcal{H}}=1} \left\| \sum_{i=1}^{\infty} a_i e_i \right\|_{L^2} \\
&\le 1. \qquad (45)
\end{aligned}$$

(ii) For the second term, we assumed $\mathcal{N}_1(\lambda) \ln n / n = o(1)$. Therefore (33) and (34) imply that, for any fixed $\delta \in (0,1)$, when n is sufficiently large, with probability at least $1 - \delta$

$$\left\| T_\lambda^{\frac{1}{2}} T_{\boldsymbol{X}\lambda}^{-1} T_\lambda^{\frac{1}{2}} \right\| \le \left\| T_\lambda^{\frac{1}{2}} T_{\boldsymbol{X}\lambda}^{-\frac{1}{2}} \right\| \cdot \left\| T_{\boldsymbol{X}\lambda}^{-\frac{1}{2}} T_\lambda^{\frac{1}{2}} \right\| \le 3. \qquad (46)$$

(iii) For the third term in (44), it can be rewritten as

$$\begin{aligned}
\left\| T_\lambda^{-\frac{1}{2}} \left(\tilde{g}_{\boldsymbol{Z}} - T_{\boldsymbol{X}\lambda} f_\lambda \right) \right\|_{\mathcal{H}} &= \left\| T_\lambda^{-\frac{1}{2}} \left[\left(\tilde{g}_{\boldsymbol{Z}} - \left(T_{\boldsymbol{X}} + \lambda + T - T \right) f_\lambda \right) \right] \right\|_{\mathcal{H}} \\
&= \left\| T_\lambda^{-\frac{1}{2}} \left[\left(\tilde{g}_{\boldsymbol{Z}} - T_{\boldsymbol{X}} f_\lambda \right) - \left(T + \lambda \right) f_\lambda + T f_\lambda \right] \right\|_{\mathcal{H}} \\
&= \left\| T_\lambda^{-\frac{1}{2}} \left[\left(\tilde{g}_{\boldsymbol{Z}} - T_{\boldsymbol{X}} f_\lambda \right) - \left(g - T f_\lambda \right) \right] \right\|_{\mathcal{H}}. \qquad (47)
\end{aligned}$$

Denote $\xi_i = \xi(\boldsymbol{x}_i) = T_\lambda^{-\frac{1}{2}} (K_{\boldsymbol{x}_i} f_\rho^*(\boldsymbol{x}_i) - T_{\boldsymbol{x}_i} f_\lambda)$. To use Bernstein inequality, we need to bound the m-th moment of $\xi(\boldsymbol{x})$:

$$\begin{aligned}
\mathbb{E} \left\| \xi(\boldsymbol{x}) \right\|_{\mathcal{H}}^m &= \mathbb{E} \left\| T_\lambda^{-\frac{1}{2}} K_{\boldsymbol{x}} (f_\rho^* - f_\lambda(\boldsymbol{x})) \right\|_{\mathcal{H}}^m \\
&\le \mathbb{E} \Big(\left\| T_\lambda^{-\frac{1}{2}} k(\boldsymbol{x}, \cdot) \right\|_{\mathcal{H}}^m \mathbb{E} \big(\left| (f_\rho^* - f_\lambda(\boldsymbol{x})) \right|^m \mid \boldsymbol{x} \big) \Big). \qquad (48)
\end{aligned}$$

Note that Lemma 39 shows that

$$\left\| T_\lambda^{-\frac{1}{2}} k(\boldsymbol{x}, \cdot) \right\|_{\mathcal{H}} \le \mathcal{N}_1(\lambda)^{\frac{1}{2}}, \quad \mu\text{-a.e. } \boldsymbol{x} \in \mathcal{X};$$

By definition of $\mathcal{M}_1(\lambda)$, we also have

$$\left\| f_\lambda - f_\rho^* \right\|_{L^\infty} \le \operatorname*{ess\,sup}_{\boldsymbol{x} \in \mathcal{X}} \left| \sum_{i=1}^{\infty} \frac{\lambda}{\lambda_i + \lambda} f_i e_i(\boldsymbol{x}) \right| = \mathcal{M}_1(\lambda). \qquad (49)$$

In addition, we have proved in Lemma 16 that

$$\mathbb{E}|(f_\lambda(\boldsymbol{x}) - f_\rho^*(\boldsymbol{x}))|^2 = \mathcal{M}_2(\lambda).$$

So we get the upper bound of (48), i.e.,

$$\begin{aligned}(48) &\leq \mathcal{N}_1(\lambda)^{\frac{m}{2}} \cdot \|f_\lambda - f_\rho^*\|_{L^\infty}^{m-2} \cdot \mathbb{E}|(f_\lambda(\boldsymbol{x}) - f_\rho^*(\boldsymbol{x}))|^2 \\ &\leq \mathcal{N}_1(\lambda)^{\frac{m}{2}} \mathcal{M}_1(\lambda)^{m-2} \mathcal{M}_2(\lambda) \\ &\leq \left(\mathcal{N}_1(\lambda)^{\frac{1}{2}} \mathcal{M}_1(\lambda)\right)^{m-2} \left(\mathcal{N}_1(\lambda)^{\frac{1}{2}} \mathcal{M}_2(\lambda)^{\frac{1}{2}}\right)^2.\end{aligned}$$

Using Lemma 38 with therein notations: $L = \mathcal{N}_1(\lambda)^{\frac{1}{2}} \mathcal{M}_1(\lambda)$ and $\sigma = \mathcal{N}_1(\lambda)^{\frac{1}{2}} \mathcal{M}_2(\lambda)^{\frac{1}{2}}$, for any fixed $\delta \in (0,1)$, with probability at least $1-\delta$, we have

$$(47) \leq 4\sqrt{2} \log \frac{2}{\delta} \left(\frac{\mathcal{N}_1(\lambda)^{\frac{1}{2}} \mathcal{M}_1(\lambda)}{n} + \frac{\mathcal{N}_1(\lambda)^{\frac{1}{2}} \mathcal{M}_2(\lambda)^{\frac{1}{2}}}{\sqrt{n}} \right). \tag{50}$$

Since we have assumed $n^{-1}\mathcal{N}_1(\lambda)^{\frac{1}{2}} \mathcal{M}_1(\lambda) = o\left(\mathcal{M}_2(\lambda)^{\frac{1}{2}}\right)$ and $\mathcal{N}_1(\lambda) \ln n / n = o(1)$, (50) further implies

$$\left\| T_\lambda^{-\frac{1}{2}} \left(\tilde{g}_{\boldsymbol{Z}} - T_{\boldsymbol{X}\lambda} f_\lambda \right) \right\|_{\mathcal{H}} = o_{\mathbb{P}}\left(\mathcal{M}_2(\lambda)^{\frac{1}{2}}\right). \tag{51}$$

Plugging (45), (46) and (51) into (44), we finish the proof.

■

Final proof of the bias term. Now we are ready to state the theorem about the bias term.

Theorem 18 *Suppose that Assumption 1, 2 and 3 hold. If the following approximation conditions hold for some $\lambda = \lambda(d,n) \to 0$:*

$$\frac{\mathcal{N}_1(\lambda)}{n} \ln n = o(1); \quad \textit{and} \quad n^{-1}\mathcal{N}_1(\lambda)^{\frac{1}{2}} \mathcal{M}_1(\lambda) = o\left(\mathcal{M}_2(\lambda)^{\frac{1}{2}}\right), \tag{52}$$

then we have

$$\mathbf{Bias}^2(\lambda) = \Theta_{\mathbb{P}}\left(\mathcal{M}_2(\lambda)\right). \tag{53}$$

Proof The triangle inequality implies that

$$\mathbf{Bias}(\lambda) = \left\| \tilde{f}_\lambda - f_\rho^* \right\|_{L^2} \geq \left\| f_\lambda - f_\rho^* \right\|_{L^2} - \left\| \tilde{f}_\lambda - f_\lambda \right\|_{L^2},$$

When $\lambda = \lambda(d,n)$ satisfies (52), Lemma 16 and Lemma 17 prove that

$$\left\| f_\lambda - f_\rho^* \right\|_{L^2} = \mathcal{M}_2(\lambda)^{\frac{1}{2}}; \quad \left\| \tilde{f}_\lambda - f_\lambda \right\|_{L^2} = o_{\mathbb{P}}(\mathcal{M}_2(\lambda)^{\frac{1}{2}}),$$

which directly prove (53). ■

A.4 Final proof of Theorem 1

Now we are ready to prove Theorem 1. Note that we have assumed in Theorem 1 that $\lambda = \lambda(d,n) \to 0$ satisfies all the conditions required in Theorem 15 and Theorem 18. Therefore, Theorem 15 and Theorem 18 show that

$$\mathbf{Var}(\lambda) = \Theta_{\mathbb{P}}\left(\frac{\sigma^2 \mathcal{N}_2(\lambda)}{n}\right); \quad \mathbf{Bias}^2(\lambda) = \Theta_{\mathbb{P}}\left(\mathcal{M}_2(\lambda)\right).$$

Recalling the bias-variance decomposition (18), we finish the proof.

■

Appendix B. Proof of inner product kernel

In this section, we aim to apply Theorem 1 to prove the results in Section 3.2. We will see that the application is nontrivial and is an important contribution of this paper.

We first introduce more necessary preliminaries in Appendix B.1, which is a preparation for subsequent calculations. Next, in order to apply Theorem 1 to get specific convergence rates, we calculate the exact convergence rates of the key quantities therein in Appendix B.2. Finally, we state the proof of Theorem 4 and Theorem 5 in turn in Appendix B.3 and B.4. We will see that there are essential differences in the proof of these two theorems.

B.1 More preliminaries about the inner product kernel on the unit sphere

Suppose that $\mathcal{X} = \mathbb{S}^d$ and μ is the uniform distribution on $\mathbb{S}^d$. Recall that in Section 3.2, we consider the inner product kernel, i.e., there exists a function $\Phi(t) : [-1,1] \to \mathbb{R}$ such that $k(\boldsymbol{x}, \boldsymbol{x}') = \Phi\left(\langle \boldsymbol{x}, \boldsymbol{x}' \rangle\right), \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{S}^d$. Then Mercer's decomposition for the inner product kernel is given in the basis of spherical harmonics:

$$k(\boldsymbol{x}, \boldsymbol{y}) = \sum_{k=0}^{\infty} \mu_k \sum_{l=1}^{N(d,k)} Y_{k,l}(\boldsymbol{x}) Y_{k,l}(\boldsymbol{y}), \tag{54}$$

where $\{Y_{k,l}\}_{l=1}^{N(d,k)}$ are spherical harmonic polynomials of degree k; μ_k are the eigenvalues with multiplicity $N(d,0) = 1$; $N(d,k) = \frac{2k+d-1}{k} \cdot \frac{(k+d-2)!}{(d-1)!(k-1)!}, k = 1, 2, \cdots$.

By known results on spherical harmonics, the eigenvalues μ_k's have the following explicit expression (Bietti and Mairal, 2019):

$$\mu_k = \frac{\omega_{d-1}}{\omega_d} \int_{-1}^{1} \Phi(t) P_k(t) \left(1 - t^2\right)^{(d-2)/2} \, \mathrm{d}t, \tag{55}$$

where P_k is the k-th Legendre polynomial in dimension $d+1$, ω_d denotes the surface of the unit sphere $\mathbb{S}^d$.

Although the above expression of $\mu_k, N(d,k)$ are complicated, Lemma 19 – 21 (mainly cited from Lu et al. 2023) give concise characterizations of μ_k and $N(d,k)$, which is sufficient for the analysis in this paper.

Lemma 19 *Let $k = k_d$ be the inner product kernel on $\mathbb{S}^d$ satisfying Assumption 1 and 4. For any fixed integer $p \geq 0$, there exist constants $\mathfrak{C}, \mathfrak{C}_1$ and $\mathfrak{C}_2$ only depending on p and $\{a_j\}_{j\leq p+1}$, such that for any $d \geq \mathfrak{C}$, we have*

$$\mathfrak{C}_1 d^{-k} \leq \mu_k \leq \mathfrak{C}_2 d^{-k}, \quad k = 0, 1, \cdots, p+1.$$

Proof From equation (22) in Ghorbani et al. (2021), for any integer $p \geq 0$, there exist constants $\mathfrak{C}$ only depending on p and $\{a_j\}_{j\leq p+1}$, such that for any $d \geq \mathfrak{C}$, we have

$$\frac{\Phi^{(k)}(0)}{d^k} \leq \mu_k \leq \frac{2\Phi^{(k)}(0)}{d^k}, \quad k = 0, 1, \cdots, p+1.$$

Note that for any $k \geq 0$, we have $a_k = \Phi^{(k)}(0)$. Therefore, letting $\mathfrak{C}_1 := \min_{k\leq p+1}\{a_k\} > 0$ and $\mathfrak{C}_2 := 2\max_{k\leq p+1}\{a_k\} < \infty$, then we finish the proof. ■

The following property of the eigenvalues $\{\mu_k\}_{k\geq 0}$ indicates that when we consider μ_p with any fixed $p \geq 0$, the subsequent eigenvalues μ_k's ($k \geq p+1$) are much smaller than μ_p. The following lemma is the same as Lemma 3.3 in Lu et al. (2023).

Lemma 20 *Let $k = k_d$ be the inner product kernel on $\mathbb{S}^d$ satisfying Assumption 1 and 4. For any fixed integer $p \geq 0$, there exist constants $\mathfrak{C}$ only depending on p and $\{a_j\}_{j\leq p+1}$, such that for any $d \geq \mathfrak{C}$, we have*

$$\mu_k \leq \frac{\mathfrak{C}_2}{\mathfrak{C}_1} d^{-1} \mu_p, \quad k = p+1, p+2, \cdots$$

where $\mathfrak{C}_1$ and $\mathfrak{C}_2$ are constants given in Lemma 19.

Lemma 21 *For an integer $k \geq 0$, denote $N(d,k)$ as the multiplicity of the eigenspace corresponding to μ_k in the Mercer's decomposition. For any fixed integer $p \geq 0$, there exist constants $\mathfrak{C}_3, \mathfrak{C}_4$ and $\mathfrak{C}$ only depending on p, such that for any $d \geq \mathfrak{C}$, we have*

$$\mathfrak{C}_3 d^k \leq N(d,k) \leq \mathfrak{C}_4 d^k, \quad k = 0, 1, \cdots, p+1. \tag{56}$$

Proof When $k = 0$, we have $N(d,0) = 1$, which satisfies (56). When $k \geq 1$, Section 1.6 in Gallier et al. (2020) shows that

$$N(d,k) = \frac{2k+d-1}{k} \cdot \frac{(k+d-2)!}{(d-1)!(k-1)!}.$$

Note that p is fixed and we consider those $k \leq p+1$, (56) follows from detailed calculations using Stirling's approximation. We refer to Lemma B.1 and Lemma D.4 in Lu et al. (2023) for more details. ■

The following lemma verifies that if we consider the inner product kernel on the unit sphere, then Assumption 3 naturally holds.

Lemma 22 *Suppose that $\mathcal{X} = \mathbb{S}^d$ and μ is the uniform distribution on $\mathbb{S}^d$. Suppose that k is an inner product kernel, then Assumption 3 holds.*

Proof Recall that we have the mercer decomposition (54). Define the sum

$$Z_{k,d}(\boldsymbol{x},\boldsymbol{y}) = \sum_{l=1}^{N(d,k)} Y_{k,l}(\boldsymbol{x})Y_{k,l}(\boldsymbol{y}).$$

Then Dai and Xu (2013, Corollary 1.2.7) shows that $Z_{k,d}(\boldsymbol{x},\boldsymbol{y})$ depends only on $\langle x, y\rangle$ and satisfies

$$|Z_{k,d}(\boldsymbol{x},\boldsymbol{y})| \le Z_{k,d}(\boldsymbol{x},\boldsymbol{x}) = N(d,k), \quad \forall x, y \in \mathbb{S}^d.$$

Therefore, we have

$$\begin{aligned}\sup_{\boldsymbol{x}\in\mathcal{X}} \sum_{i=1}^{\infty} \left(\frac{\lambda_i}{\lambda_i+\lambda}\right)^2 e_i^2(\boldsymbol{x}) &= \sup_{\boldsymbol{x}\in\mathcal{X}} \sum_{k=1}^{\infty}\sum_{l=1}^{N(d,k)} \left(\frac{\mu_k}{\mu_k+\lambda}\right)^2 Y_{k,l}^2(\boldsymbol{x}) = \sum_{k=1}^{\infty} \left(\frac{\mu_k}{\mu_k+\lambda}\right)^2 \sup_{\boldsymbol{x}\in\mathcal{X}} Z_{k,d}(\boldsymbol{x},\boldsymbol{x}) \\ &= \sum_{k=1}^{\infty} \left(\frac{\mu_k}{\mu_k+\lambda}\right)^2 N(d,k) = \sum_{k=1}^{\infty}\sum_{l=1}^{N(d,k)} \left(\frac{\mu_k}{\mu_k+\lambda}\right)^2 \\ &= \mathcal{N}_2(\lambda).\end{aligned}$$

The other equation in Assumption 3 can be proved similarly. ■

B.2 Calculations of some key quantities

Based on the information of the eigenvalues in the last subsection, this subsection determines the exact convergence rates of the quantities appeared in Theorem 1. These rates will finally determine the convergence rates in Theorem 4 and Theorem 5. Note that we assume that d grows with n in Theorem 4 and Theorem 5 .

Lemma 23 *Consider $\mathcal{X} = \mathbb{S}^d$ and the marginal distribution μ to be the uniform distribution. Let $k = k_d$ be the inner product kernel on $\mathbb{S}^d$ satisfying Assumption 1 and 4. By choosing $\lambda = d^{-l}$ for some $l > 0$, we have:*

$$\mathcal{N}_1(\lambda) = \Theta\left(\lambda^{-1}\right); \tag{57}$$

If $p \le l \le p+1$ for some $p \in \{0, 1, 2\cdots\}$, we have

$$\mathcal{N}_2(\lambda) = \Theta\left(d^p + \lambda^{-2}d^{-(p+1)}\right). \tag{58}$$

The notation Θ involves constants only depending on κ and p.

Proof If $p \le l \le p+1$ for some $p \in \{0, 1, 2\cdots\}$, Lemma 19 and Lemma 21 show that there exist constants $\mathfrak{C}, \mathfrak{C}_1, \mathfrak{C}_2, \mathfrak{C}_3$ and $\mathfrak{C}_4$ only depending on p (recall that we ignore the dependence on $\{a_j\}_{j=0}^{\infty}$), such that for any $d \ge \mathfrak{C}$, we have

$$\mathfrak{C}_2^{-1}\mu_{p+1} \le \lambda \le \mathfrak{C}_1^{-1}\mu_p;$$

and for $k = 0, 1, \cdots, p+1$,

$$\mathfrak{C}_1 d^{-k} \leq \mu_k \leq \mathfrak{C}_2 d^{-k}; \quad \mathfrak{C}_3 d^k \leq N(d,k) \leq \mathfrak{C}_4 d^k.$$

We first prove (57). On the one hand, for any $d \geq \mathfrak{C}$, we have

$$\begin{aligned}
\mathcal{N}_1(\lambda) &= \sum_{k=0}^{p} \frac{\mu_k}{\mu_k + \lambda} N(d,k) + \sum_{k=p+1}^{\infty} \frac{\mu_k}{\mu_k + \lambda} N(d,k) \\
&\leq \sum_{k=0}^{p} \frac{\mu_k}{\mu_k} N(d,k) + \sum_{k=p+1}^{\infty} \frac{\mu_k}{\lambda} N(d,k) \\
&\leq p\mathfrak{C}_4 d^p + \lambda^{-1} \sum_{k=p+1}^{\infty} \mu_k N(d,k) \\
&\leq p\mathfrak{C}_4 d^p + \lambda^{-1}\kappa^2 \\
&\lesssim \lambda^{-1}, \qquad (59)
\end{aligned}$$

where we use the fact that $\sum_{k=p+1}^{\infty} \mu_k N(d,k) \leq \sum_{i=1}^{\infty} \lambda_i \leq \sup_{\boldsymbol{x} \in \mathcal{X}} k(x,x) \leq \kappa^2$ for the third inequality. On the other hand, for any $d \geq \mathfrak{C}$, we have

$$\begin{aligned}
\mathcal{N}_1(\lambda) &= \sum_{k=0}^{p} \frac{\mu_k}{\mu_k + \lambda} N(d,k) + \sum_{k=p+1}^{\infty} \frac{\mu_k}{\mu_k + \lambda} N(d,k) \\
&\geq \frac{\mu_p}{\mu_p + \lambda} N(d,p) + \frac{\mu_{p+1}}{\mu_{p+1} + \lambda} N(d,p+1) \\
&\geq \frac{\mu_p}{\mu_p + \mathfrak{C}_1^{-1}\mu_p} N(d,p) + \frac{\mu_{p+1}}{\mathfrak{C}_2\lambda + \lambda} N(d,p+1) \\
&\geq \frac{\mathfrak{C}_3}{1 + \mathfrak{C}_1^{-1}} d^p + \frac{\mathfrak{C}_1\mathfrak{C}_3}{\mathfrak{C}_2 + 1} \lambda^{-1} \\
&\gtrsim \lambda^{-1}. \qquad (60)
\end{aligned}$$

Combining (59) and (60), we finish the proof of (57).

Now we begin to prove (58). On the one hand, for any $d \geq \mathfrak{C}$ we have

$$\begin{aligned}
\mathcal{N}_2(\lambda) &= \sum_{k=0}^{p} \left(\frac{\mu_k}{\mu_k + \lambda} \right)^2 N(d,k) + \sum_{k=p+1}^{\infty} \left(\frac{\mu_k}{\mu_k + \lambda} \right)^2 N(d,k) \\
&\leq \sum_{k=0}^{p} \left(\frac{\mu_k}{\mu_k} \right)^2 N(d,k) + \sum_{k=p+1}^{\infty} \left(\frac{\mu_k}{\lambda} \right)^2 N(d,k) \\
&\leq p\mathfrak{C}_4 d^p + \lambda^{-2} \sum_{k=p+1}^{\infty} \mu_k^2 N(d,k) \\
&\leq p\mathfrak{C}_4 d^p + \lambda^{-2}\mu_{p+1} \sum_{k=p+1}^{\infty} \mu_k N(d,k) \\
&\leq p\mathfrak{C}_4 d^p + \lambda^{-2}\mathfrak{C}_2 d^{-(p+1)}\kappa^2. \qquad (61)
\end{aligned}$$

Note that for the forth equation, we use the fact that $\mu_k \le \mu_{p+1}, \forall k \ge p+1$ (when d is sufficiently large), which can be proved by Lemma 20. On the other hand, for any $d \ge \mathfrak{C}$, we have

$$\begin{aligned}\mathcal{N}_2(\lambda) &= \sum_{k=0}^{p}\left(\frac{\mu_k}{\mu_k+\lambda}\right)^2 N(d,k) + \sum_{k=p+1}^{\infty}\left(\frac{\mu_k}{\mu_k+\lambda}\right)^2 N(d,k)\\ &\ge \left(\frac{\mu_k}{\mu_p+\mathfrak{C}_1^{-1}\mu_p}\right)^2 N(d,p) + \left(\frac{\mu_{p+1}}{\mathfrak{C}_2\lambda+\lambda}\right)^2 N(d,p+1)\\ &\ge \frac{\mathfrak{C}_3}{\left(1+\mathfrak{C}_1^{-1}\right)^2} d^p + \lambda^{-2}\left(\frac{\mathfrak{C}_1}{\mathfrak{C}_2+1}\right)^2 \mathfrak{C}_3 d^{-(p+1)}. \end{aligned} \tag{62}$$

Combining (61) and (62), we prove that $\mathcal{N}_2(\lambda) = \Theta\left(d^p + \lambda^{-2}d^{-(p+1)}\right)$.

■

Before stating lemmas about $\mathcal{M}_1(\lambda), \mathcal{M}_2(\lambda)$, we first introduce the following useful lemma.

Lemma 24 *Consider $\mathcal{X} = \mathbb{S}^d$ and the marginal distribution μ to be the uniform distribution. Let $k = k_d$ be the inner product kernel on $\mathbb{S}^d$ satisfying Assumption 1 and 4. Further suppose that Assumption 5 holds for some $s > 0$. Suppose that α, β are two real numbers such that*

$$s + \alpha - \beta \le 0; \quad s + \alpha \ge 0. \tag{63}$$

Then by choosing $\lambda = d^{-l}$ for some $0 < l < \gamma$, if $p \le l \le p+1$ for some $p \in \{0, 1, 2\cdots\}$, we have

$$\mathcal{M}_{\alpha,\beta}(\lambda) := \sum_{i=1}^{\infty} \frac{\lambda_i^\alpha}{(\lambda_i+\lambda)^\beta} f_i^2 = \Theta\left(d^{-(s+\alpha-\beta)p} + \lambda^{-\beta} d^{-(p+1)(s+\alpha)}\right).$$

The notation Θ involves constants only depending on s, p, c_0 and R_γ, where c_0 and R_γ are the constants from Assumption 5.

Proof Similar as the proof of Lemma 23, if $p \le l \le p+1$ for some $p \in \{0, 1, 2\cdots\}$, there exist constants $\mathfrak{C}, \mathfrak{C}_1, \mathfrak{C}_2, \mathfrak{C}_3$ and $\mathfrak{C}_4$ only depending on p, such that for any $d \ge \mathfrak{C}$, we have

$$\mathfrak{C}_2^{-1}\mu_{p+1} \le \lambda \le \mathfrak{C}_1^{-1}\mu_p;$$

and for $k = 0, 1, \cdots, p+1$,

$$\mathfrak{C}_1 d^{-k} \le \mu_k \le \mathfrak{C}_2 d^{-k}; \quad \mathfrak{C}_3 d^k \le N(d,k) \le \mathfrak{C}_4 d^k.$$

On the one hand, since (63) holds, for any $d \geq \mathfrak{C}$, we have

$$\begin{aligned}
\mathcal{M}_{\alpha,\beta}(\lambda) &= \sum_{k=0}^{\infty} \left(\frac{\mu_k^{s+\alpha}}{(\mu_k+\lambda)^{\beta}} \sum_{i \in \mathcal{I}_k} \mu_k^{-s} f_i^2 \right) \\
&\leq \sum_{k=0}^{p} \frac{\mu_k^{s+\alpha}}{(\mu_k+\lambda)^{\beta}} R_\gamma^2 + \sum_{k=p+1}^{\infty} \frac{\mu_k^{s+\alpha}}{(\mu_k+\lambda)^{\beta}} \sum_{i \in \mathcal{I}_k} \mu_k^{-s} f_i^2 \\
&\leq \sum_{k=0}^{p} \mu_k^{s+\alpha-\beta} R_\gamma^2 + \sum_{k=p+1}^{\infty} \frac{\mu_k^{s+\alpha}}{\lambda^{\beta}} \sum_{i \in \mathcal{I}_k} \mu_k^{-s} f_i^2 \\
&\leq p \mu_p^{s+\alpha-\beta} R_\gamma^2 + \lambda^{-\beta} \mu_{p+1}^{s+\alpha} \sum_{k=p+1}^{\infty} \sum_{i \in \mathcal{I}_k} \mu_k^{-s} f_i^2 \\
&\leq p R_\gamma^2 \mathfrak{C}_2^{(s+\alpha-\beta)} d^{-(s+\alpha-\beta)p} + \lambda^{-\beta} \mathfrak{C}_2^{s+\alpha} d^{-(p+1)(s+\alpha)} R_\gamma^2 \\
&\lesssim d^{-(s+\alpha-\beta)p} + \lambda^{-\beta} d^{-(p+1)(s+\alpha)}. \qquad (64)
\end{aligned}$$

Note that Assumption 5 (a) implies $\sum\limits_{k=0}^{\infty} \sum\limits_{i \in \mathcal{I}_k} \mu_k^{-s} f_i^2 = \sum\limits_{i=1}^{\infty} \lambda_i^{-s} f_i^2 \leq R_\gamma^2$; We also use the fact that $\mu_k \leq \mu_{p+1}, \forall k \geq p+1$, which can be proved by Lemma 20.
On the other hand, for any $d \geq \mathfrak{C}$, we have

$$\begin{aligned}
\mathcal{M}_{\alpha,\beta}(\lambda) &= \sum_{k=0}^{\infty} \left(\frac{\mu_k^{s+\alpha}}{(\mu_k+\lambda)^{\beta}} \sum_{i \in \mathcal{I}_k} \mu_k^{-s} f_i^2 \right) \\
&\geq \frac{\mu_p^{s+\alpha}}{(\mu_p+\lambda)^{\beta}} \sum_{i \in \mathcal{I}_p} \mu_p^{-s} f_i^2 + \frac{\mu_{p+1}^{s+\alpha}}{(\mu_{p+1}+\lambda)^{\beta}} \sum_{i \in \mathcal{I}_{d,p+1}} \mu_{p+1}^{-s} f_i^2 \\
&\geq \frac{\mu_p^{s+\alpha}}{\left(\mu_p + \mathfrak{C}_1^{-1}\mu_p\right)^{\beta}} \sum_{i \in \mathcal{I}_p} \mu_p^{-s} f_i^2 + \frac{\mu_{p+1}^{s+\alpha}}{(\mathfrak{C}_2\lambda+\lambda)^{\beta}} \sum_{i \in \mathcal{I}_{p+1}} \mu_{p+1}^{-s} f_i^2 \\
&\geq \frac{\mathfrak{C}_3^{s+\alpha-\beta}}{\left(1+\mathfrak{C}_1^{-1}\right)^{\beta}} d^{-(s+\alpha-\beta)p} c_0 + \lambda^{-\beta} \frac{\mathfrak{C}_1^{s+\alpha}}{(\mathfrak{C}_2+1)^{\beta}} d^{-(p+1)(s+\alpha)} c_0 \\
&\gtrsim d^{-(s+\alpha-\beta)p} + \lambda^{-\beta} d^{-(p+1)(s+\alpha)}. \qquad (65)
\end{aligned}$$

We use Assumption 5 (b), i.e., $\sum\limits_{i \in \mathcal{I}_p} \mu_p^{-s} f_i^2 \geq c_0$ and $\sum\limits_{i \in \mathcal{I}_{p+1}} \mu_{p+1}^{-s} f_i^2 \geq c_0$, to obtain the lower bound. Combining (64) and (65), we finish the proof.

■

Lemma 25 *Consider $\mathcal{X} = \mathbb{S}^d$ and the marginal distribution μ to be the uniform distribution. Let $k = k_d$ be the inner product kernel on $\mathbb{S}^d$ satisfying Assumption 1 and 4. Further suppose that Assumption 5 holds for some $s > 0$. Define $\tilde{s} = \min\{s, 2\}$. By choosing $\lambda = d^{-l}$ for some $0 < l < \gamma$, we have: if $p \leq l \leq p+1$ for some $p \in \{0, 1, 2, \cdots\}$, we have*

$$\mathcal{M}_2(\lambda) = \Theta\left(\lambda^2 d^{(2-\tilde{s})p} + d^{-(p+1)\tilde{s}}\right). \qquad (66)$$

The notation Θ involves constants only depending on s, p, c_0 and R_γ, where c_0 and R_γ are the constants from Assumption 5.

Proof When $0 < s \le 2$, $\mathcal{M}_2(\lambda)$ can be viewed as $\lambda^2 \mathcal{M}_{\alpha,\beta}(\lambda)$ in Lemma 24 with $\alpha = 0, \beta = 2$. The conditions (63) are satisfied and Lemma 24 shows that

$$\mathcal{M}_2(\lambda) = \Theta\left(\lambda^2 d^{(2-s)p} + d^{-(p+1)s}\right). \tag{67}$$

When $s > 2$, without loss of generality, we can assume $\lambda_i \le 1, \forall i$ and $\lambda \le 1$. Recall we have proved in Lemma 19 that μ_0 remains as a constant as $d \to \infty$. On the one hand, Assumption 5 (b) also implies $f_1^2 \ge c_0$, which further implies

$$\mathcal{M}_2(\lambda) = \lambda^2 \sum_{i=1}^{\infty} \frac{f_i^2}{(\lambda_i + \lambda)^2} \ge \frac{1}{4}\lambda^2 \sum_{i=1}^{\infty} f_i^2 \ge \frac{1}{4}\lambda^2 c_0. \tag{68}$$

On the other hand, since $s > 2$, we have

$$\begin{aligned}
\mathcal{M}_2(\lambda) &= \lambda^2 \sum_{k=0}^{\infty} \left(\frac{\mu_k^s}{(\mu_k + \lambda)^2} \sum_{i \in \mathcal{I}_k} \mu_k^{-s} f_i^2 \right) \\
&\le \lambda^2 \left(\sup_{k \ge 0} \frac{\mu_k^s}{(\mu_k + \lambda)^2} \cdot \sum_{k=0}^{\infty} \sum_{i \in \mathcal{I}_k} \mu_k^{-s} f_i^2 \right) \\
&\le \lambda^2 \cdot \sup_{k \ge 0} \frac{\mu_k^s}{(\mu_k + \lambda)^2} \cdot R_\gamma^2 \\
&\le \lambda^2 R_\gamma^2.
\end{aligned} \tag{69}$$

Further note that since $s > 2$ and $p \le l \le p + 1$, (68) and (69) implies

$$\mathcal{M}_2(\lambda) = \Theta\left(\lambda^2\right) = \Theta\left(\lambda^2 d^{(2-\tilde{s})p} + d^{-(p+1)\tilde{s}}\right).$$

We finish the proof. ■

The following lemma applies for those $s \ge 1$, which gives an upper bound of $\mathcal{M}_1(\lambda)$.

Lemma 26 *Consider $\mathcal{X} = \mathbb{S}^d$ and the marginal distribution μ to be the uniform distribution. Let $k = k_d$ be the inner product kernel on $\mathbb{S}^d$ satisfying Assumption 1 and 4. Further suppose that Assumption 5 holds for some $s \ge 1$. Define $\tilde{s} = \min\{s, 2\}$. By choosing $\lambda = d^{-l}$ for some $0 < l < \gamma$, we have: if $p \le l \le p + 1$ for some $p \in \{0, 1, 2 \cdots\}$, we have*

$$\mathcal{M}_1(\lambda) = O\left(\lambda^{\frac{1}{2}} d^{\frac{(2-\tilde{s})p}{2}} + d^{-\frac{(\tilde{s}-1)(p+1)}{2}}\right). \tag{70}$$

The notation O involves constants only depending on s, κ, p and R_γ, where R_γ is the constant from Assumption 5.

Proof First, Cauchy-Schwarz inequality shows that

$$\begin{aligned}\mathcal{M}_1(\lambda)^2 &= \operatorname*{ess\,sup}_{\boldsymbol{x}\in\mathcal{X}} \left|\sum_{i=1}^{\infty}\left(\frac{\lambda}{\lambda_i+\lambda} f_i e_i(\boldsymbol{x})\right)\right|^2 \\ &\le \left(\sum_{i=1}^{\infty}\frac{\lambda^2\lambda_i^{-1}}{\lambda_i+\lambda} f_i^2\right)\cdot \operatorname*{ess\,sup}_{\boldsymbol{x}\in\mathcal{X}} \sum_{i=1}^{\infty}\left(\frac{\lambda_i}{\lambda_i+\lambda} e_i(\boldsymbol{x})^2\right) \\ &\le \left(\sum_{i=1}^{\infty}\frac{\lambda^2\lambda_i^{-1}}{\lambda_i+\lambda} f_i^2\right)\cdot \sum_{i=1}^{\infty}\left(\frac{\lambda_i}{\lambda_i+\lambda}\right) \\ &:= \mathcal{Q}_1(\lambda)\cdot\mathcal{N}_1(\lambda), \end{aligned} \tag{71}$$

where we use (6) in Assumption 3 for the second inequality.

When $1 \le s \le 2$, $\mathcal{Q}_1(\lambda)$ defined above can be viewed as $\lambda^2\mathcal{M}_{\alpha,\beta}(\lambda)$ in Lemma 24 with $\alpha = -1, \beta = 1$. In addition, the conditions (63) are satisfied, thus Lemma 24 shows that

$$\mathcal{Q}_1(\lambda) = \Theta\left(\lambda^2 d^{(2-s)p} + \lambda d^{-(s-1)(p+1)}\right). \tag{72}$$

Since we assume the kernel to be bounded in Assumption 1, we can assume $\lambda_i \le 1, \forall i$ and $\lambda \le 1$ without loss of generality. When $s > 2$, on the one hand, Assumption 5 (b) also implies

$$\mathcal{Q}_1(\lambda) = \lambda^2\sum_{i=1}^{\infty}\frac{f_i^2}{(\lambda_i+\lambda)\lambda_i} \ge \frac{1}{2}\lambda^2\sum_{i=1}^{\infty} f_i^2 \ge \frac{1}{2}\lambda^2 c_0. \tag{73}$$

On the other hand, since $s > 2$, we have

$$\begin{aligned}\mathcal{Q}_1(\lambda) &= \lambda^2\sum_{k=0}^{\infty}\left(\frac{\mu_k^{s-1}}{\mu_k+\lambda}\sum_{i\in\mathcal{I}_k}\mu_k^{-s} f_i^2\right) \\ &\le \lambda^2\left(\sup_{k\ge 0}\frac{\mu_k^{s-1}}{\mu_k+\lambda}\cdot\sum_{k=0}^{\infty}\sum_{i\in\mathcal{I}_k}\mu_k^{-s} f_i^2\right) \\ &\le \lambda^2\cdot\frac{\mu_k^{s-1}}{\mu_k+\lambda}\cdot R_\gamma^2 \\ &\le \lambda^2 R_\gamma^2. \end{aligned} \tag{74}$$

Further note that since $s > 2$ and $p \le l \le p+1$, (73) and (74) implies

$$\mathcal{Q}_1(\lambda) = \Theta\left(\lambda^2\right) = \Theta\left(\lambda^2 d^{(2-\tilde{s})p} + \lambda d^{-(\tilde{s}-1)(p+1)}\right).$$

Therefore, we have $\mathcal{Q}_1(\lambda) = \Theta\left(\lambda^2 d^{(2-\tilde{s})p} + \lambda d^{-(\tilde{s}-1)(p+1)}\right)$ for any $s > 0$.

Further recalling that Lemma 23 proves $\mathcal{N}_1(\lambda) = \Theta\left(\lambda^{-1}\right)$, use (71) and we finish the proof.

■

When $0 < s < 1$, the following lemma gives an upper bound of $\|f_\lambda\|_{L^\infty}$.

Lemma 27 *Consider $\mathcal{X} = \mathbb{S}^d$ and the marginal distribution μ to be the uniform distribution. Let $k = k_d$ be the inner product kernel on $\mathbb{S}^d$ satisfying Assumption 1 and 4. Further suppose that Assumption 5 holds for some $0 < s < 1$. Recall the definition of f_λ in (17). By choosing $\lambda = d^{-l}$ for some $0 < l < \gamma$, we have: if $p \le l \le p+1$ for some $p \in \{0, 1, 2 \cdots\}$, we have*

$$\|f_\lambda\|_{L^\infty} = O\left(d^{\frac{(1-s)p}{2}} + \lambda^{-1} d^{-\frac{(1+s)(p+1)}{2}}\right). \tag{75}$$

The notation O involves constants only depending on s, κ, p, and R_γ, where R_γ is the constant from Assumption 5.

Proof We need the following fact: For any $f \in \mathcal{H}$, since Assumption 1 holds, we have

$$\begin{aligned}\|f_\lambda\|_{L^\infty} &\le \sup_{\boldsymbol{x}\in\mathcal{X}} |f(\boldsymbol{x})| = \sup_{\boldsymbol{x}\in\mathcal{X}} \langle f(\cdot), k(\boldsymbol{x}, \cdot)\rangle_{\mathcal{H}} \\ &\le \sup_{\boldsymbol{x}\in\mathcal{X}} \|k(\boldsymbol{x}, \cdot)\|_{\mathcal{H}} \cdot \|f\|_{\mathcal{H}} \\ &\le \kappa^2 \|f\|_{\mathcal{H}}.\end{aligned}$$

Therefore, by definition and notice that $f_\lambda \in \mathcal{H}$, we have

$$\|f_\lambda\|_{L^\infty}^2 \le \kappa^4 \|f_\lambda\|_{\mathcal{H}}^2 = \kappa^4 \sum_{i=1}^{\infty} \frac{\lambda_i}{(\lambda_i + \lambda)^2} f_i^2 := \mathcal{Q}_2(\lambda).$$

Since $0 < s < 1$, $\mathcal{Q}_2(\lambda)$ can be viewed as $\kappa^4 \mathcal{M}_{\alpha,\beta}(\lambda)$ in Lemma 24 with $\alpha = 1, \beta = 2$ and the conditions (63) are satisfied. Thus Lemma 24 shows that

$$\mathcal{Q}_2(\lambda) = \Theta\left(\lambda^2 d^{(1-s)p} + \lambda^{-s} d^{-(1+s)(p+1)}\right).$$

Taking the square root, we finish the proof. ■

B.3 Proof of Theorem 4

In the last subsection, we have calculated the exact convergence rates of $\mathcal{N}_1(\lambda), \mathcal{N}_2(\lambda)$, $\mathcal{M}_1(\lambda)$, and $\mathcal{M}_2(\lambda)$ when $\lambda = d^{-l}$ for some $0 < l < \gamma$. Note that we have proved in Lemma 22 that Assumption 3 naturally holds for inner product kernel on the unit sphere. Now we are ready to apply Theorem 1 to prove Theorem 4. The proof mainly consists of 3 steps:

(1) For specific range of $\gamma > 0$, we use Lemma 23 and Lemma 25 to derive the scale of λ_{balance} or l_{balance} such that $\mathcal{N}_2(\lambda)/n$ and $\mathcal{M}_2(\lambda)$ are balanced.

(2) We check that the conditions (7) required in Theorem 1 are satisfied for $\lambda \gtrsim \lambda_{\text{balance}}$ (or $\lambda = d^{-l}, l \le l_{\text{balance}}$).

(3) Using the monotonicity of $\mathbf{Var}(\lambda)$ with respect to λ, we demonstrate that $\lambda = \lambda_{\text{balance}}$ is the best choice of regularization parameter, i.e., the generalization error of KRR estimator is the smallest when $\lambda = \lambda_{\text{balance}}$. That is to say, the convergence rate of the generalization error can not be faster than the rate when choosing $\lambda = \lambda_{\text{balance}}$.

Note that we expect $\mathbf{Bias}^2(\lambda) = \Theta_{\mathbb{P}}(\mathcal{M}_2(\lambda))$ and $\mathbf{Var}(\lambda) = \Theta_{\mathbb{P}}(\mathcal{N}_2(\lambda)/n)$. Step 1 actually indicates the regularization such that the bias and variance are balanced. Together with Theorem 15 and 18, Step 2 further verifies that $\mathbf{Bias}^2(\lambda) = \Theta_{\mathbb{P}}(\mathcal{M}_2(\lambda))$ and $\mathbf{Var}(\lambda) = \Theta_{\mathbb{P}}(\mathcal{N}_2(\lambda)/n)$ indeed hold for those $\lambda \gtrsim \lambda_{\text{balance}}$. Thus they are indeed balanced under the choice of $\lambda = \lambda_{\text{balance}}$.

Final proof of Theorem 4. In the following of the proof, we omit the dependence of constants on $s, \sigma, \gamma, c_0, \kappa, c_1$ and c_2.
Step 1: Note that we assume $s \geq 1$ in this theorem and $\lambda = d^{-l}, 0 < l < \gamma$. For specific range of γ, we discuss the range of l_{balance}. Recall that we define $\tilde{s} = \min\{s, 2\}$.

- When $l \in (p, p + \frac{1}{2}]$ for some integer $p \geq 0$, Lemma 23 and Lemma 25 show that

$$\frac{\mathcal{N}_2(\lambda)}{n} \asymp d^{p-\gamma}; \quad \mathcal{M}_2(\lambda) \asymp d^{-2l+(2-\tilde{s})p},$$

  thus we have

$$l_{\text{balance}} = \frac{\gamma + p - p\tilde{s}}{2}.$$

  Further, letting $l_{\text{balance}} = \frac{\gamma+p-p\tilde{s}}{2} \in (p, p + \frac{1}{2}]$, we have

$$\gamma \in (p + p\tilde{s}, p + p\tilde{s} + 1].$$

- When $l \in (p + \frac{1}{2}, p + \frac{\tilde{s}}{2}]$, Lemma 23 and Lemma 25 show that

$$\frac{\mathcal{N}_2(\lambda)}{n} \asymp d^{2l-p-1-\gamma}; \quad \mathcal{M}_2(\lambda) \asymp d^{-2l+(2-\tilde{s})p},$$

  thus we have

$$l_{\text{balance}} = \frac{\gamma + 3p - p\tilde{s} + 1}{4}.$$

  Further, letting $l_{\text{balance}} = \frac{\gamma+3p-p\tilde{s}+1}{4} \in (p + \frac{1}{2}, p + \frac{\tilde{s}}{2}]$, we have

$$\gamma \in (p + p\tilde{s} + 1, p + p\tilde{s} + 2\tilde{s} - 1].$$

- When $l \in (p + \frac{\tilde{s}}{2}, p + 1]$, Lemma 23 and Lemma 25 show that

$$\frac{\mathcal{N}_2(\lambda)}{n} \asymp d^{2l-p-1-\gamma}; \quad \mathcal{M}_2(\lambda) \asymp d^{-(p+1)\tilde{s}},$$

  thus we have

$$l_{\text{balance}} = \frac{\gamma + (p+1)(1-\tilde{s})}{2}.$$

  Further, letting $l_{\text{balance}} \in (p + \frac{\tilde{s}}{2}, p + 1]$, we have

$$\gamma \in (p + p\tilde{s} + 2\tilde{s} - 1, (p+1) + (p+1)\tilde{s}].$$

Step 2: In order to apply Theorem 15 and Theorem 18 so that we know the exact convergence rates of $\mathbf{Var}(\lambda_{\text{balance}})$ and $\mathbf{Bias}^2(\lambda_{\text{balance}})$, we first check the approximation conditions (38) and (43), or equivalently conditions (7), hold for $l = l_{\text{balance}}$. Recall that we have calculated the convergence rates of $\mathcal{N}_1(\lambda)$ and $\mathcal{M}_1(\lambda)$ in Lemma 23 and Lemma 26.

- When $\gamma \in (p + p\tilde{s}, p + p\tilde{s} + 1]$: recall that $l_{\text{balance}} = \frac{\gamma+p-p\tilde{s}}{2} \in (p, p + \frac{1}{2}]$.

  (i) The first condition in (7) is equivalent to

  $$\frac{\gamma + p - p\tilde{s}}{2} < \gamma \iff \gamma > p - p\tilde{s},$$

  which naturally holds for all $\gamma \in (p + p\tilde{s}, p + p\tilde{s} + 1]$.

  (ii) The second condition in (7) is equivalent to

  $$d^{-\gamma} \cdot d^{\gamma+p-p\tilde{s}} \cdot \gamma \ln d \ll d^p \iff p - p\tilde{s} < p,$$

  which naturally holds for all $\gamma \in (p + p\tilde{s}, p + p\tilde{s} + 1]$ and $p \neq 0$. When $p = 0$, we actually need to choose $\lambda_{\text{balance}} = d^{-l_{\text{balance}}} \cdot \ln d$ and the second condition will hold.

  (iii) The third condition in (7) is equivalent to

  $$d^{-\gamma} \cdot d^{\frac{\gamma+p-p\tilde{s}}{4}} \cdot \left[d^{-\frac{\gamma+p-p\tilde{s}}{4} + \frac{(2-\tilde{s})p}{2}} + d^{-\frac{(\tilde{s}-1)(p+1)}{2}} \right] \ll d^{-\frac{1}{2}(\gamma+p-p\tilde{s}) + \frac{(2-\tilde{s})p}{2}}$$

  $\iff$

  $$\gamma > p - p\tilde{s}; \quad \gamma > p - 3p\tilde{s} - 2\tilde{s} + 2,$$

  which naturally holds for all $\gamma \in (p + p\tilde{s}, p + p\tilde{s} + 1]$ and $p \neq 0$. In addition, one can also check that the third condition in (7) holds when $p = 0$ and $\lambda_{\text{balance}} = d^{-l_{\text{balance}}} \cdot \ln d$.

- When $\gamma \in (p + p\tilde{s} + 1, p + p\tilde{s} + 2\tilde{s} - 1]$: recall that $l_{\text{balance}} = \frac{\gamma+3p-p\tilde{s}+1}{4} \in (p + \frac{1}{2}, p + \frac{\tilde{s}}{2}]$.

  (i) The first condition in (7) is equivalent to

  $$\frac{\gamma + 3p - p\tilde{s} + 1}{4} < \gamma \iff \gamma > p - \frac{p\tilde{s}}{3} + \frac{1}{3},$$

  which naturally holds for all $\gamma \in (p + p\tilde{s} + 1, p + p\tilde{s} + 2\tilde{s} - 1]$.

  (ii) The second condition in (7) is equivalent to

  $$d^{-\gamma} \cdot d^{\frac{\gamma+3p-p\tilde{s}+1}{2}} \cdot \gamma \ln d \ll d^{\frac{\gamma+3p-p\tilde{s}+1}{2} - p - 1} \iff \gamma > p + 1,$$

  which naturally holds for all $\gamma \in (p + p\tilde{s} + 1, p + p\tilde{s} + 2\tilde{s} - 1]$.

  (iii) The third condition in (7) is equivalent to

  $$d^{-\gamma} \cdot d^{\frac{\gamma+p-p\tilde{s}}{8}} \cdot \left[d^{-\frac{\gamma+p-p\tilde{s}}{8} + \frac{(2-\tilde{s})p}{2}} + d^{-\frac{(\tilde{s}-1)(p+1)}{2}} \right] \ll d^{-\frac{1}{4}(\gamma+p-p\tilde{s}) + \frac{(2-\tilde{s})p}{2}}$$

  $\iff$

  $$\gamma > p - \frac{p\tilde{s}}{3} + \frac{1}{3}; \quad \gamma > p - \frac{3p\tilde{s}}{5} - \frac{4\tilde{s}}{5} + \frac{7}{5},$$

  which naturally holds for all $\gamma \in (p + p\tilde{s} + 1, p + p\tilde{s} + 2\tilde{s} - 1]$.

- When $\gamma \in (p+p\tilde{s}+2\tilde{s}-1, (p+1)+(p+1)\tilde{s}]$: recall that $l_{\text{balance}} = \frac{\gamma+(p+1)(1-\tilde{s})}{2} \in (p+\frac{\tilde{s}}{2}, p+1$.

  (i) The first condition in (7) is equivalent to

$$\frac{\gamma+(p+1)(1-\tilde{s})}{2} < \gamma \iff \gamma > p - p\tilde{s} + 1 - \tilde{s},$$

  which naturally holds for all $\gamma \in (p+p\tilde{s}+2\tilde{s}-1, (p+1)+(p+1)\tilde{s}]$.

  (ii) The second condition in (7) is equivalent to

$$d^{-\gamma} \cdot d^{\gamma+(p+1)(1-\tilde{s})} \cdot \gamma \ln d \ll d^{\gamma+(p+1)(1-\tilde{s})-p-1} \iff \gamma > p+1,$$

  which naturally holds for all $\gamma \in (p+p\tilde{s}+2\tilde{s}-1, (p+1)+(p+1)\tilde{s}]$.

  (iii) The third condition in (7) is equivalent to

$$d^{-\gamma} \cdot d^{\frac{\gamma+(p+1)(1-\tilde{s})}{4}} \cdot \left[d^{-\frac{\gamma+(p+1)(1-\tilde{s})}{2} + \frac{(2-\tilde{s})p}{2}} + d^{-\frac{(\tilde{s}-1)(p+1)}{2}} \right] \ll d^{-\frac{(p+1)\tilde{s}}{2}}$$

$$\iff$$

$$\gamma > p + \frac{\tilde{s}}{2}; \quad \gamma > p - \frac{p\tilde{s}}{3} + \frac{\tilde{s}}{3} + \frac{1}{3},$$

  which naturally holds for all $\gamma \in (p+p\tilde{s}+2\tilde{s}-1, (p+1)+(p+1)\tilde{s}]$.

Up to now, we have verified conditions (7) for $l = l_{\text{balance}}$. Furthermore, simple calculation shows that the order of

$$\frac{\mathcal{N}_1(\lambda)}{n}; \quad n^{-1}\mathcal{N}_1(\lambda)^2\mathcal{N}_2(\lambda)^{-1}; \quad n^{-1}\mathcal{N}_1(\lambda)^{\frac{1}{2}}\mathcal{M}_1(\lambda)\mathcal{M}_2(\lambda)^{-\frac{1}{2}} \tag{76}$$

are all non-decreasing with respect to l, where we choose $\lambda = d^{-l}$. Therefore, the above results indicate that conditions (7) holds for all $l \le l_{\text{balance}}$.

Step 3: In step 2, on the one hand, we prove that by choosing $\lambda = \lambda_{\text{balance}}$,

$$\mathbb{E}\left[\left\| \hat{f}_{\lambda_{\text{balance}}} - f_\rho^* \right\|_{L^2}^2 \;\middle|\; \boldsymbol{X} \right] = \Theta_{\mathbb{P}}\left(\frac{\sigma^2\mathcal{N}_2(\lambda_{\text{balance}})}{n} + \mathcal{M}_2(\lambda_{\text{balance}}) \right). \tag{77}$$

On the other hand, we also prove that by choosing $\lambda \gtrsim \lambda_{\text{balance}}$,

$$\mathbb{E}\left[\left\| \hat{f}_{\lambda_{\text{balance}}} - f_\rho^* \right\|_{L^2}^2 \;\middle|\; \boldsymbol{X} \right] = \Omega_{\mathbb{P}}\left(\frac{\sigma^2\mathcal{N}_2(\lambda_{\text{balance}})}{n} + \mathcal{M}_2(\lambda_{\text{balance}}) \right). \tag{78}$$

In the following, we handle those $\lambda \lesssim \lambda_{\text{balance}}$. Recall that we have shown in the proof of Lemma 11 that

$$\mathbf{Var}(\lambda) = \frac{\sigma^2}{n^2} \int_{\mathcal{X}} \mathbb{K}(\boldsymbol{x}, \boldsymbol{X})(\boldsymbol{K}+\lambda)^{-2}\mathbb{K}(\boldsymbol{X}, \boldsymbol{x}) \, \mathrm{d}\mu(\boldsymbol{x}).$$

One simple but critical observation from Li et al. (2023a) is that

$$(\boldsymbol{K}+\lambda_1)^{-2} \succeq (\boldsymbol{K}+\lambda_2)^{-2}, \quad \text{if} \quad \lambda_1 \le \lambda_2,$$

where $\succeq$ represents the partial order of positive semi-definite matrices, and thus for $\lambda_1 \leq \lambda_2 > 0$, we have

$$\mathbf{Var}(\lambda_1) \geq \mathbf{Var}(\lambda_2).$$

Also note that by the definition of l_{balance}, we actually have

$$\mathbb{E}\left[\left\|\hat{f}_{\lambda_{\text{balance}}} - f_\rho^*\right\|_{L^2}^2 \;\middle|\; \boldsymbol{X}\right] = \Theta_{\mathbb{P}}\left(\mathbf{Var}(\lambda_{\text{balance}})\right).$$

Therefore, for those $\lambda \lesssim \lambda_{\text{balance}}$, we have

$$\mathbb{E}\left[\left\|\hat{f}_{\lambda} - f_\rho^*\right\|_{L^2}^2 \;\middle|\; \boldsymbol{X}\right] \geq \mathbf{Var}(\lambda) \geq \mathbf{Var}(\lambda_{\text{balance}}) \asymp \mathbb{E}\left[\left\|\hat{f}_{\lambda_{\text{balance}}} - f_\rho^*\right\|_{L^2}^2 \;\middle|\; \boldsymbol{X}\right]. \tag{79}$$

To sum up, (77), (78) and (79) show that by choosing $\lambda = \lambda_{\text{balance}}$ as in step 1, we obtain the convergence rates of KRR estimator under the best regularization. Using Lemma 25 to calculate the rate of $\mathcal{M}_2(\lambda_{\text{balance}})$, we finish the proof.

■

B.4 Proof of Theorem 5

Recall that in the proof of Theorem 18, we use $\mathcal{M}_1(\lambda)$ to bound $\left\|f_\lambda - f_\rho^*\right\|_{L^\infty}$ and show that $\left\|\tilde{f}_\lambda - f_\lambda\right\|_{L^2} = o_{\mathbb{P}}(\mathcal{M}_2(\lambda)^{\frac{1}{2}})$. Unfortunately, the calculation of $\mathcal{M}_1(\lambda)$ in Lemma 26 only holds for $s \geq 1$ and it could be infinite when $s < 1$. Extensive literature then assume $\left\|f_\rho^*\right\|_{L^\infty}$ to be bounded and use $\|f_\lambda\|_{L^\infty} + \left\|f_\rho^*\right\|_{L^\infty}$ to bound $\left\|f_\lambda - f_\rho^*\right\|_{L^\infty}$. When the dimension d is fixed, Zhang et al. (2024) first use a truncation method together with the L^q-embedding property of $[\mathcal{H}]^s$ to remove the boundedness assumption when $s < 1$. We will see in the proof of Theorem 5 that this technique still works in the large-dimensional setting.

To be specific, when we further assume $f_\rho^* \in [\mathcal{H}]^s$, we have the following Theorem which is a refined version of Lemma 17.

Lemma 28 *Suppose that Assumption 1, 2 and 3 hold. Further suppose that Assumption 5 holds for some $0 < s < 1$. If the following approximation conditions hold for some $\lambda = \lambda(d, n) \to 0$:*

$$\frac{\mathcal{N}_1(\lambda)}{n} \ln n = o(1); \quad n^{-1}\mathcal{N}_1(\lambda)^{\frac{1}{2}} \|f_\lambda\|_{L^\infty} = o\left(\mathcal{M}_2(\lambda)^{\frac{1}{2}}\right); \tag{80}$$

and there exists $\varepsilon > 0$, such that

$$n^{-1}\mathcal{N}_1(\lambda)^{\frac{1}{2}} n^{\frac{1-s}{2}+\varepsilon} = o\left(\mathcal{M}_2(\lambda)^{\frac{1}{2}}\right), \tag{81}$$

then we have

$$\left\|\tilde{f}_\lambda - f_\lambda\right\|_{L^2} = o_{\mathbb{P}}(\mathcal{M}_2(\lambda)^{\frac{1}{2}}),$$

where the notation $o_{\mathbb{P}}$ involves constants only depending on s and κ.

Proof Recall the decomposition (44) in the proof of Lemma 17. The first two terms in (44) can be handled without any difference. Our goal here is to prove the following equation and we will finish the proof:

$$\left\| T_\lambda^{-\frac{1}{2}} \left(\tilde{g}_{\boldsymbol{Z}} - T_{\boldsymbol{X}\lambda} f_\lambda \right) \right\|_{\mathcal{H}} = o_{\mathbb{P}} \left(\mathcal{M}_2(\lambda)^{\frac{1}{2}} \right). \tag{82}$$

Similarly as (47), we rewrite the left hand side of (82) as

$$\left\| T_\lambda^{-\frac{1}{2}} \left(\tilde{g}_{\boldsymbol{Z}} - T_{\boldsymbol{X}\lambda} f_\lambda \right) \right\|_{\mathcal{H}} = \left\| T_\lambda^{-\frac{1}{2}} \left[\left(\tilde{g}_{\boldsymbol{Z}} - T_{\boldsymbol{X}} f_\lambda \right) - \left(g - T f_\lambda \right) \right] \right\|_{\mathcal{H}}. \tag{83}$$

Denote $\xi_i = \xi(\boldsymbol{x}_i) = T_\lambda^{-\frac{1}{2}} (K_{\boldsymbol{x}_i} f_\rho^*(\boldsymbol{x}_i) - T_{\boldsymbol{x}_i} f_\lambda)$. Further consider the subset $\Omega_1 = \{\boldsymbol{x} \in \mathcal{X} : |f_\rho^*(\boldsymbol{x})| \le t\}$ and $\Omega_2 = \mathcal{X} \backslash \Omega_1$, where t will be chosen appropriately later. Decompose ξ_i as $\xi_i I_{\boldsymbol{x}_i \in \Omega_1} + \xi_i I_{\boldsymbol{x}_i \in \Omega_2}$ and we have the following decomposition of (83):

$$\begin{aligned} \left\| \frac{1}{n} \sum_{i=1}^n \xi_i - \mathbb{E}\xi_x \right\|_{\mathcal{H}} &\le \left\| \frac{1}{n} \sum_{i=1}^n \xi_i I_{\boldsymbol{x}_i \in \Omega_1} - \mathbb{E}\xi_{\boldsymbol{x}} I_{\boldsymbol{x} \in \Omega_1} \right\|_{\mathcal{H}} + \| \frac{1}{n} \sum_{i=1}^n \xi_i I_{\boldsymbol{x}_i \in \Omega_2} \|_{\mathcal{H}} + \| \mathbb{E}\xi_{\boldsymbol{x}} I_{\boldsymbol{x} \in \Omega_2} \|_{\mathcal{H}} \\ &:= \mathrm{I} + \mathrm{II} + \mathrm{III}. \end{aligned} \tag{84}$$

Next we choose $t = n^{\frac{1-s}{2} + \varepsilon_t}, q = \frac{2}{1-s} - \varepsilon_q$ such that

$$\varepsilon_t < \varepsilon; \quad \text{and} \quad \frac{1-s}{2} + \varepsilon_t > 1 / \left(\frac{2}{1-s} - \varepsilon_q \right), \tag{85}$$

where ε is given in (81). Then we can bound the three terms in (84) as follows:

(i) For the first term in (84), denoted as I, notice that

$$\left\| \left(f_\lambda - f_\rho^* \right) I_{\boldsymbol{x}_i \in \Omega_1} \right\|_{L^\infty} \le \| f_\lambda \|_{L^\infty} + n^{\frac{1-s}{2} + \varepsilon_t}.$$

Imitating the procedure (iii) in the proof of Lemma 17 and using (80), (81), we have

$$\mathrm{I} = o_{\mathbb{P}} \left(\mathcal{M}_2(\lambda)^{\frac{1}{2}} \right). \tag{86}$$

(ii) For the second term in (84), denoted as II. Since $q = \frac{2}{1-s} - \varepsilon_q < \frac{2}{1-s}$, Lemma 44 shows that,

$$[\mathcal{H}]^s \hookrightarrow L^q(\mathcal{X}, \mu),$$

with embedding norm less than a constant $C_{s,\kappa}$. Then Assumption 5 (a) implies that there exists $0 < C_q < \infty$ only depending on γ, s and κ such that $\| f_\rho^* \|_{L^q(\mathcal{X},\mu)} \le C_q$. Using the Markov inequality, we have

$$P(\boldsymbol{x} \in \Omega_2) = P\Big(|f_\rho^*(\boldsymbol{x})| > t \Big) \le \frac{\mathbb{E}|f_\rho^*(\boldsymbol{x})|^q}{t^q} \le \frac{(C_q)^q}{t^q}.$$

Further, since (85) guarantees $t^q \gg n$, we have

$$\begin{aligned}\tau_n := P\left(\mathrm{II} \geq \mathcal{M}_2(\lambda)^{\frac{1}{2}}\right) &\leq P\Big(\ \exists \boldsymbol{x}_i \text{ s.t. } \boldsymbol{x}_i \in \Omega_2,\Big) = 1 - P\Big(\boldsymbol{x}_i \notin \Omega_2, \forall \boldsymbol{x}_i, i = 1, 2, \cdots, n\Big) \\ &= 1 - P\Big(\boldsymbol{x} \notin \Omega_2\Big)^n \\ &= 1 - P\Big(|f_\rho^*(\boldsymbol{x})| \leq t\Big)^n \\ &\leq 1 - \Big(1 - \frac{(C_q)^q}{t^q}\Big)^n \to 0. \end{aligned} \tag{87}$$

(iii) For the third term in (84), denoted as III. Since Lemma 39 implies that $\|T_\lambda^{-\frac{1}{2}} k(\boldsymbol{x}, \cdot)\|_{\mathcal{H}} \leq \mathcal{N}_1(\lambda)^{\frac{1}{2}}$, μ-a.e. $\boldsymbol{x} \in \mathcal{X}$, so

$$\begin{aligned}\mathrm{III} &\leq \mathbb{E}\|\xi_{\boldsymbol{x}} I_{\boldsymbol{x}\in\Omega_2}\|_{\mathcal{H}} \leq \mathbb{E}\Big[\|T_\lambda^{-\frac{1}{2}} k(\boldsymbol{x}, \cdot)\|_{\mathcal{H}} \cdot \big|\big(f_\rho^* - f_\lambda(\boldsymbol{x})\big) I_{\boldsymbol{x}\in\Omega_2}\big|\Big] \\ &\leq \mathcal{N}_1(\lambda)^{\frac{1}{2}} \mathbb{E}\big|\big(f_\rho^* - f_\lambda(\boldsymbol{x})\big) I_{\boldsymbol{x}\in\Omega_2}\big| \\ &\leq \mathcal{N}_1(\lambda)^{\frac{1}{2}} \left\|f_\rho^* - f_\lambda\right\|_{L^2}^{\frac{1}{2}} \cdot P\left(\boldsymbol{x} \in \Omega_2\right)^{\frac{1}{2}} \\ &\leq \mathcal{N}_1(\lambda)^{\frac{1}{2}} \mathcal{M}_2(\lambda)^{\frac{1}{2}} t^{-\frac{q}{2}}, \end{aligned} \tag{88}$$

where we use Cauchy-Schwarz inequality for the third inequality and (16) for the forth inequality. Recalling that the choices of t, q satisfy $t^{-q} \ll n^{-1}$ and we have assumed $\mathcal{N}_1(\lambda) \ln n / n = o(1)$, we have

$$\mathrm{III} = o\left(\mathcal{M}_2(\lambda)^{\frac{1}{2}}\right). \tag{89}$$

Plugging (86), (87) and (89) into (84), we finish the proof.

■

Based on Lemma 28 and Lemma 16, we have the following theorem about the exact rate of bias term when $0 < s < 1$.

Theorem 29 *Suppose that Assumption 1, 2 and 3 hold. Further suppose that Assumption 5 holds for some $0 < s < 1$. If the following approximation conditions hold for some $\lambda = \lambda(d, n) \to 0$:*

$$\frac{\mathcal{N}_1(\lambda)}{n} \ln n = o(1); \quad n^{-1} \mathcal{N}_1(\lambda)^{\frac{1}{2}} \|f_\lambda\|_{L^\infty} = o\left(\mathcal{M}_2(\lambda)^{\frac{1}{2}}\right); \tag{90}$$

and there exists $\varepsilon > 0$, such that

$$n^{-1} \mathcal{N}_1(\lambda)^{\frac{1}{2}} n^{\frac{1-s}{2}+\varepsilon} = o\left(\mathcal{M}_2(\lambda)^{\frac{1}{2}}\right), \tag{91}$$

then we have

$$\mathbf{Bias}^2(\lambda) = \Theta_{\mathbb{P}}\left(\mathcal{M}_2(\lambda)\right), \tag{92}$$

where the notation $\Theta_{\mathbb{P}}$ involves constants only depending on s and κ.

Proof The triangle inequality implies that

$$\mathbf{Bias}(\lambda) = \left\|\tilde{f}_\lambda - f_\rho^*\right\|_{L^2} \geq \left\|f_\lambda - f_\rho^*\right\|_{L^2} - \left\|\tilde{f}_\lambda - f_\lambda\right\|_{L^2},$$

When $\lambda = \lambda(d, n)$ satisfies (90) and (91), Lemma 16 and Lemma 28 prove that

$$\left\|f_\lambda - f_\rho^*\right\|_{L^2} = \mathcal{M}_2(\lambda)^{\frac{1}{2}}; \quad \left\|\tilde{f}_\lambda - f_\lambda\right\|_{L^2} = o_{\mathbb{P}}(\mathcal{M}_2(\lambda)^{\frac{1}{2}}),$$

which directly prove (92). ■

Now we are ready to prove Theorem 5. Since we do not claim that the regularization choice in Theorem 5 is the best, we only need the first two steps in the proof of Theorem 4.

Final proof of Theorem 5. In the following of the proof, we omit the dependence of constants on $s, \sigma, \gamma, c_0, \kappa, c_1$ and c_2.
Step 1: Note that we assume $0 < s < 1$ in this theorem and $\lambda = d^{-l}, 0 < l < \gamma$. For specific range of γ, we discuss the range of l_{balance}.

- When $l \in (p, p + \frac{s}{2}]$ for some integer $p \geq 0$, Lemma 23 and Lemma 25 show that

  $$\frac{\mathcal{N}_2(\lambda)}{n} \asymp d^{p-\gamma}; \quad \mathcal{M}_2(\lambda) \asymp d^{-2l+(2-s)p},$$

  thus we have

  $$l_{\text{balance}} = \frac{\gamma + p - ps}{2}.$$

  Further, letting $l_{\text{balance}} = \frac{\gamma+p-ps}{2} \in (p, p + \frac{s}{2}]$, we have

  $$\gamma \in (p + ps, p + ps + s].$$

- When $l \in (p + \frac{s}{2}, p + \frac{1}{2}]$, Lemma 23 and Lemma 25 show that

  $$\frac{\mathcal{N}_2(\lambda)}{n} \asymp d^{p-\gamma}; \quad \mathcal{M}_2(\lambda) \asymp d^{-(p+1)s},$$

  thus the above two terms are equal if and only if

  $$\gamma = p + ps + s.$$

- When $l \in (p + \frac{1}{2}, p + 1]$, Lemma 23 and Lemma 25 show that

  $$\frac{\mathcal{N}_2(\lambda)}{n} \asymp d^{2l-p-1-\gamma}; \quad \mathcal{M}_2(\lambda) \asymp d^{-(p+1)s},$$

  thus we have

  $$l_{\text{balance}} = \frac{\gamma + (p+1)(1-s)}{2}.$$

  Further, letting $l_{\text{balance}} \in (p + \frac{s}{2}, p + 1]$, we have

  $$\gamma \in (p + ps + s, (p+1) + (p+1)s].$$

Note that the present result is different from the result of the Step 1 in the proof of Theorem (4). There are only two intervals of γ, i.e.,

$$\gamma \in (p+ps, p+ps+s]\,; \quad \text{and} \quad \gamma \in (p+ps+s, (p+1)+(p+1)s]\,.$$

It is worth mentioning that in the second interval of γ, we can actually choose $\lambda = d^{-l}, \forall l \in \left[p+\frac{s}{2}, l_{\text{balance}}\right]$ and we have

$$\frac{\mathcal{N}_2(\lambda)}{n} \lesssim \mathcal{M}_2(\lambda); \quad \mathcal{M}_2(\lambda) = \mathcal{M}_2(\lambda_{\text{balance}}).$$

That is to say, we can choose smaller l and the rate of $\mathcal{N}_2(\lambda)/n + \mathcal{M}_2(\lambda)$ will remain unchanged. We have shown in (76) and the discussion below it that the approximation conditions are easier to satisfied for smaller l. Therefore, in the following of the proof, we define

$$l_{\text{opt}} = p + \frac{s}{2}, \quad \text{when} \quad \gamma \in (p+ps+s, (p+1)+(p+1)s]\,,$$

and verify the approximation conditions for $\lambda_{\text{opt}} = d^{-l_{\text{opt}}}$. For consistency of notation, we also define

$$l_{\text{opt}} = \frac{\gamma + p - ps}{2}, \quad \text{when} \quad \gamma \in (p+ps, p+ps+s]\,.$$

Step 2: In order to apply Theorem 15 and Theorem 29 so that we know the exact convergence rates of $\mathbf{Var}(\lambda_{\text{opt}})$ and $\mathbf{Bias}^2(\lambda_{\text{opt}})$, we first check the approximation conditions (38), (90) and (91) hold for $l = l_{\text{opt}}$. We first list all the approximation conditions below:

$$\begin{aligned} &\frac{\mathcal{N}_1(\lambda)}{n}\ln n = o(1); \quad n^{-1}\mathcal{N}_1(\lambda)^2 \ln n = o(\mathcal{N}_2(\lambda)); \\ &n^{-1}\mathcal{N}_1(\lambda)^{\frac{1}{2}} \|f_\lambda\|_{L^\infty} = o\left(\mathcal{M}_2(\lambda)^{\frac{1}{2}}\right); n^{-1}\mathcal{N}_1(\lambda)^{\frac{1}{2}} n^{\frac{1-s}{2}+\varepsilon} = o\left(\mathcal{M}_2(\lambda)^{\frac{1}{2}}\right). \end{aligned} \tag{93}$$

Recall that we have calculated the convergence rates of $\mathcal{N}_1(\lambda)$ and $\|f_\lambda\|_{L^\infty}$ in Lemma 23 and Lemma 27.

- When $\gamma \in (p+ps, p+ps+s]$: recall that $l_{\text{opt}} = \frac{\gamma+p-ps}{2} \in (p, p+\frac{s}{2}]$.

  (i) The first condition in (93) is equivalent to

  $$\frac{\gamma+p-ps}{2} < \gamma \iff \gamma > p - ps,$$

  which naturally holds for all $\gamma \in (p+ps, p+ps+s]$.

  (ii) The second condition in (93) is equivalent to

  $$d^{-\gamma} \cdot d^{\gamma+p-ps} \cdot \gamma \ln d \ll d^p \iff p - ps < p,$$

  which naturally holds for all $\gamma \in (p+ps, p+ps+s]$ and $p \neq 0$. When $p = 0$, we actually need to choose $\lambda_{\text{opt}} = d^{-l_{\text{opt}}} \cdot \ln d$ and the second condition will hold.

(iii) The third condition in (93) is equivalent to

$$d^{-\gamma}\cdot d^{\frac{\gamma+p-ps}{4}}\cdot\left[d^{\frac{p}{2}-\frac{ps}{2}}+d^{\frac{\gamma+p-ps}{2}-\frac{(1+s)(p+1)}{2}}\right]\ll d^{-\frac{1}{2}(\gamma+p-ps)+\frac{(2-s)p}{2}}$$

$\Longleftrightarrow$

$$\gamma>p-3ps;\quad \gamma<p+5ps+2s+2,$$

which naturally holds for all $\gamma\in(p+ps,p+ps+s]$ and $p\neq 0$. In addition, one can also check that the third condition in (93) holds when $p=0$ and $\lambda_{\text{opt}}=d^{-l_{\text{opt}}}\cdot\ln d$.
(iv) The forth condition in (93) is equivalent to

$$d^{-\gamma}\cdot d^{\frac{\gamma+p-ps}{4}}\cdot d^{\frac{\gamma}{2}-\frac{\gamma s}{2}}\ll d^{-\frac{1}{2}(\gamma+p-ps)+\frac{(2-s)p}{2}}$$

$\Longleftrightarrow$

$$(1-2s)\gamma<p+ps. \tag{94}$$

If $\frac{1}{2}<s<1$, (94) naturally holds for all $\gamma\in(p+ps,p+ps+s]$ and $p\neq 0$. In addition, one can also check that the forth condition in (93) holds when $p=0$ and $\lambda_{\text{opt}}=d^{-l_{\text{opt}}}\cdot\ln d$.
If $0<s\leq\frac{1}{2}$, (94) only holds for $\gamma\in(p+ps,p+ps+s]$ and $p\neq 0$. That is to say, we can not verify (94) holds for

$$\gamma\in(0,s],\quad 0<s<\frac{1}{2}. \tag{95}$$

- When $\gamma\in(p+ps+s,(p+1)+(p+1)s]$: recall that $l_{\text{opt}}=p+\frac{s}{2}$.
(i) The first condition in (93) is equivalent to

$$p+\frac{s}{2}<\gamma,$$

which naturally holds for all $\gamma\in(p+ps+s,(p+1)+(p+1)s]$.

(ii) The second condition in (93) is equivalent to

$$d^{-\gamma}\cdot d^{2p+s}\cdot\gamma\ln d\ll d^{p}\iff\gamma>p+s,$$

which naturally holds for all $\gamma\in(p+ps+s,(p+1)+(p+1)s]$.

(iii) The third condition in (93) is equivalent to

$$d^{-\gamma}\cdot d^{\frac{p}{2}+\frac{s}{4}}\cdot\left[d^{\frac{p}{2}-\frac{ps}{2}}+d^{p+\frac{s}{2}-\frac{(1+s)(p+1)}{2}}\right]\ll d^{-\frac{(p+1)s}{2}}$$

$\Longleftrightarrow$

$$\gamma>p+\frac{3s}{4};\quad \gamma>p+\frac{3s}{4}-\frac{1}{2},$$

which naturally holds for all $\gamma\in(p+ps+s,(p+1)+(p+1)s]$.
(iv) The forth condition in (93) is equivalent to

$$d^{-\gamma}\cdot d^{\frac{p}{2}+\frac{s}{4}}\cdot d^{\frac{\gamma}{2}-\frac{\gamma s}{2}}\ll d^{-\frac{(p+1)s}{2}}$$

$$\Longleftrightarrow$$

$$\gamma > \frac{2p+3s+2ps}{2(s+1)}. \tag{96}$$

If $1/2 < s < 1$, (96) naturally holds for all $\gamma \in (p+ps+s, (p+1)+(p+1)s]$ and $p \neq 0$. In addition, one can also check that the forth condition in (93) holds when $p = 0$ and $\lambda_{\text{opt}} = d^{-l_{\text{opt}}} \cdot \ln d$.
If $0 < s \leq 1/2$, (96) only holds for $\gamma \in (p+ps+s, (p+1)+(p+1)s]$ and $p \neq 0$. That is to say, we can not verify (94) holds for

$$\gamma \in \left(0, \frac{3s}{2(s+1)}\right], \quad 0 < s < \frac{1}{2}. \tag{97}$$

Up to now, we have verified the approximation conditions (93) for

$$\forall \gamma > 0, \quad \text{if} \ \ \frac{1}{2} < s < 1;$$

and

$$\forall \gamma > \frac{3s}{2(s+1)}, \quad \text{if} \ \ 0 < s \leq \frac{1}{2}.$$

Using Lemma 25 to calculate the rate of $\mathcal{M}_2(\lambda_{\text{opt}})$, we finish the proof.

■

Appendix C. Proof of Minimax lower bound

C.1 More preliminaries about minimax lower bound

Let's first introduce several concepts about minimax lower bound which can be frequently found in related literature Yang and Barron (1999); Lu et al. (2023), etc..

Suppose that $(\mathcal{Z}, d)$ is a topological space with a compatible loss function d, which is a mapping from $\mathcal{Z} \times \mathcal{Z}$ to $\mathbb{R}_{\geq 0}$ with $d(f, f) = 0$ and $d(f, f') > 0$ for $f \neq f'$. We call such a loss function a *distance*. We introduce the packing entropy and covering entropy below:

Definition 30 (Packing entropy) *A finite set $N_\varepsilon \subset \mathcal{Z}$ is said to be an ε-packing set in $\mathcal{Z}$ with separation $\varepsilon > 0$, if for any $f, f' \in N_\varepsilon, f \neq f'$, we have $d(f, f') > \varepsilon$. The logarithm of the maximum cardinality of ε-packing set is called the ε-packing entropy of $\mathcal{Z}$ with distance d and is denoted by $M_d(\varepsilon, \mathcal{Z})$.*

Definition 31 (Covering entropy) *A set $G_\varepsilon \subset \mathcal{Z}$ is said to be an ε-net for $\mathcal{Z}$ if for any $\tilde{f} \in \mathcal{Z}$, there exists an $f_0 \in G_\varepsilon$ such that $d(\tilde{f}, f_0) \leq \varepsilon$. The logarithm of the minimum cardinality of ε-net is called the ε-covering entropy of $\mathcal{Z}$ with distance d and is denoted by $V_d(\varepsilon, \mathcal{Z})$.*

Let $\mathcal{B} = \left\{f \in [\mathcal{H}]^s, \ \|f\|_{[\mathcal{H}]^s} \leq R_\gamma\right\}$, where R_γ is the constant from Assumption 5. Without loss of generality, we can consider $\mathcal{B}$ to be the unit ball in $[\mathcal{H}]^s$. Let $M_2(\varepsilon, \mathcal{B})$ be the

ε-packing entropy of $(\mathcal{B}, d^2 = \|\cdot\|_{L^2}^2)$ and $V_2(\varepsilon, \mathcal{B})$ be the ε-covering entropy of $(\mathcal{B}, d^2 = \|\cdot\|_{L^2}^2)$. Recalling that μ is the marginal distribution on $\mathcal{X}$, we further define

$$\mathcal{D} = \left\{ \rho_f \;\middle|\; \text{joint distribution of } (y, \boldsymbol{x}) \text{ where } \boldsymbol{x} \sim \mu, y = f(\boldsymbol{x}) + \epsilon, \epsilon \sim N(0, \sigma^2), f \in \mathcal{B} \right\},$$

and let $V_K(\varepsilon, \mathcal{D})$ be the ε-covering entropy of $(\mathcal{D}, d^2 = \text{ KL divergence })$. It is easy to see that $\mathcal{D}$ is a subset of $\mathcal{P}$ which is defined in Theorem 7, i.e., $\mathcal{D} \subset \mathcal{P}$.

The following lemmas give useful characterizations of $M_2(\varepsilon, \mathcal{B}), V_2(\varepsilon, \mathcal{B})$ and $V_K(\varepsilon, \mathcal{D})$. We refer to Lemma A.5, Lemma A.7 and Lemma A.8 in Lu et al. (2023) for their proofs.

Lemma 32 *For any $\varepsilon > 0$, we have $M_2(2\varepsilon, \mathcal{B}) \le V_2(\varepsilon, \mathcal{B}) \le M_2(\varepsilon, \mathcal{B})$.*

Lemma 33 $V_2(\varepsilon, \mathcal{B}) = V_K\left(\frac{\varepsilon}{\sqrt{2}\sigma}, \mathcal{D}\right)$.

Lemma 34 *Let $\{\lambda_j\}_{j=1}^{\infty}$ be the eigenvalues of $\mathcal{H}$. For any $\varepsilon > 0$, let $K(\varepsilon) = \frac{1}{2} \sum_{j:\lambda_j^s > \varepsilon^2} \ln\left(\lambda_j^s / \varepsilon^2\right)$. We have*

$$V_2(6\varepsilon, \mathcal{B}) \le K(\varepsilon) \le V_2(\varepsilon, \mathcal{B}).$$

The following important lemma is a modification of Theorem 1 and Corollary 1 in Yang and Barron (1999). We refer to Lemma 4.1 in Lu et al. (2023) for the proof.

Lemma 35 *Let $\mathfrak{c} \in (0, 1)$ be a constant only depending on c_1, c_2, and γ, where c_1, c_2 are the constants given in Theorem 7. For any $0 < \tilde{\varepsilon}_1, \tilde{\varepsilon}_2 < \infty$ only depending on n, d, $\{\lambda_j\}$, c_1, c_2, and γ and satisfying*

$$\frac{V_K(\tilde{\varepsilon}_2, \mathcal{D}) + n\tilde{\varepsilon}_2^2 + \ln 2}{V_2(\tilde{\varepsilon}_1, \mathcal{B})} \le \mathfrak{c},$$

we have

$$\min_{\hat{f}} \max_{\rho_{f^*} \in \mathcal{D}} \mathbb{E}_{(\boldsymbol{X}, \mathbf{y}) \sim \rho_{f^*}^{\otimes n}} \left\| \hat{f} - f^* \right\|_{L^2}^2 \ge \frac{1 - \mathfrak{c}}{4} \tilde{\varepsilon}_1^2.$$

C.2 Proof of Theorem 7

Now we are ready to use the lemmas in the last subsection to prove Theorem 7. The proof is divided into two parts, dealing with the two cases of the interval in which γ falls into.

Proof of Theorem 7 (i). In this case, we have assumed $\gamma \in (p + ps, p + ps + s]$ for some integer $p \ge 0$. Let $\tilde{\varepsilon}_2^2 = C_2 d^{-(\gamma - p)}$, where we will choose the constant C_2 later. Note that we have $\gamma - p \in (ps, (p+1)s]$. Lemma 19 implies that we can choose C_2 only depending on p (ignoring the dependence on $\{a_j\}_{j=0}^{\infty}$) such that for any $d \ge \mathfrak{C}$ ($\mathfrak{C}$ is a constant only depending on ε, s and p), we have

$$\mu_{p+1}^s < \tilde{\varepsilon}_2^2 < \mu_p^s.$$

Next we can choose $\tilde{\varepsilon}_1^2 = d^{-(\gamma - p + \varepsilon)}$, where ε can be any positive real number. Since $\gamma - p + \varepsilon > ps$, when $d \ge \mathfrak{C}$, where $\mathfrak{C}$ is a constant only depending on ε, s and p, we have

$$\tilde{\varepsilon}_1^2 < \mu_p^s.$$

Therefore, using Lemma 34 and Lemma 19, for any $d \geq \mathfrak{C}$, we have

$$\begin{aligned} V_2\left(\tilde{\varepsilon}_1, \mathcal{B}\right) \geq K(\tilde{\varepsilon}_1) &\geq \frac{1}{2} N(d,p) \ln\left(\frac{\mu_p^s}{\tilde{\varepsilon}_1^2}\right) \\ &\geq \frac{1}{2} N(d,p) \ln\left(\frac{\mathfrak{C}_1 d^{-ps}}{d^{-(\gamma-p+\varepsilon)}}\right) \\ &= \frac{1}{2} N(d,p) \left(\ln \mathfrak{C}_1 + (\gamma - p + \varepsilon - ps) \ln d\right). \end{aligned} \tag{98}$$

In addition, using Lemma 19 and Lemma 20, we have the following claim.

Claim 1 *Suppose that $\gamma \in (p+ps, p+ps+s]$ for some integer $p \geq 0$. Let $\tilde{\varepsilon}_2^2$ be defined as above. For any $\varepsilon_0 > 0$, there exists a sufficiently large constant $\mathfrak{C}$ only depending on s, p and ε_0, such that for any $d \geq \mathfrak{C}$, we have*

$$K\left(\sqrt{2}\sigma\tilde{\varepsilon}_2/6\right) \leq (1+\varepsilon_0)\frac{1}{2} N(d,p) \ln\left(\frac{18\mu_p^s}{\sigma^2\tilde{\varepsilon}_2^2}\right).$$

Therefore, for any $d \geq \mathfrak{C}$, where $\mathfrak{C}$ is a constant only depending on s and p, we have

$$\begin{aligned} V_K\left(\tilde{\varepsilon}_2, \mathcal{D}\right) = V_2\left(\sqrt{2}\sigma\tilde{\varepsilon}_2, \mathcal{B}\right) &\leq K\left(\frac{\sqrt{2}\sigma\tilde{\varepsilon}_2}{6}\right) \\ &\leq (1+\varepsilon_0)\frac{1}{2} N(d,p) \ln\left(\frac{18\mu_p^s}{\sigma^2\tilde{\varepsilon}_2^2}\right) \\ &\leq (1+\varepsilon_0)\frac{1}{2} N(d,p) \ln\left(\frac{18\mathfrak{C}_2 d^{-ps}}{\sigma^2 C_2 d^{-(\gamma-p)}}\right) \\ &= (1+\varepsilon_0)\frac{1}{2} N(d,p) \left(\ln\frac{18\mathfrak{C}_2}{\sigma^2 C_2} + (\gamma - p - ps)\ln d\right), \end{aligned} \tag{99}$$

where we use Lemma 33 and Lemma 34 for the first line and use Lemma 19 for the third line.

Using (98) and (99), also recalling that we assume $c_1 d^\gamma \leq n \leq c_2 d^\gamma$, we have

$$\frac{V_K(\tilde{\varepsilon}_2, \mathcal{D}) + n\tilde{\varepsilon}_2^2 + \ln 2}{V_2(\tilde{\varepsilon}_1, \mathcal{B})} \leq \frac{(1+\varepsilon_0)\frac{1}{2}N(d,p)\left(\ln\frac{18\mathfrak{C}_2}{\sigma^2 C_2} + (\gamma-p-ps)\ln d\right) + c_2 d^\gamma \cdot C_2 d^{-(\gamma-p)} + \ln 2}{\frac{1}{2}N(d,p)\left(\ln\mathfrak{C}_1 + (\gamma-p+\varepsilon-ps)\ln d\right)}. \tag{100}$$

Recalling that Lemma 21 shows $\mathfrak{C}_3 d^p \leq N(d,p) \leq \mathfrak{C}_4 d^p$ when $d \geq \mathfrak{C}$, the dominant terms in (100) are:

$$\frac{\frac{1}{2}(1+\varepsilon_0)(\gamma-p-ps)N(d,p)\ln d}{\frac{1}{2}(\gamma-p+\varepsilon-ps)N(d,p)\ln d}.$$

Therefore, for any $\varepsilon > 0$, we can choose ε_0 small enough such that

$$\frac{V_K(\tilde{\varepsilon}_2, \mathcal{D}) + n\tilde{\varepsilon}_2^2 + \ln 2}{V_2(\tilde{\varepsilon}_1, \mathcal{B})} \leq (100) := \mathfrak{c} < 1.$$

Then using Lemma 35, we have

$$\min_{\hat{f}} \max_{\rho_{f^*} \in \mathcal{D}} \mathbb{E}_{(\boldsymbol{X},\mathbf{y})\sim\rho_{f^*}^{\otimes n}} \left\| \hat{f} - f^* \right\|_{L^2}^2 \geq \frac{1-\mathfrak{c}}{4}\tilde{\varepsilon}_1^2 = \frac{1-\mathfrak{c}}{4} d^{-(\gamma-p-\varepsilon)}.$$

Further recalling that $\mathcal{D} \subset \mathcal{P}$, we have

$$\min_{\hat{f}} \max_{\rho \in \mathcal{P}} \mathbb{E}_{(\boldsymbol{X},\mathbf{y})\sim\rho^{\otimes n}} \left\| \hat{f} - f_\rho^* \right\|_{L^2}^2 \geq \min_{\hat{f}} \max_{\rho_{f^*} \in \mathcal{D}} \mathbb{E}_{(\boldsymbol{X},\mathbf{y})\sim\rho_{f^*}^{\otimes n}} \left\| \hat{f} - f^* \right\|_{L^2}^2 \geq \frac{1-\mathfrak{c}}{4} d^{-(\gamma-p-\varepsilon)}.$$

We finish the proof of Theorem 7 (i).

■

Proof of Theorem 7 (ii). In this case, we have assumed $\gamma \in (p+ps+s, (p+1)+(p+1)s]$ for some integer $p \geq 0$. Let $\tilde{\varepsilon}_2^2 = C_2 d^{-(p+1)s} \ln d$, where we will choose the constant C_2 later. Then Lemma 19 implies that there exists a constant $\mathfrak{C}$ only depending on s and p such that for any $d \geq \mathfrak{C}$, we have

$$\mu_{p+1}^s < \tilde{\varepsilon}_2^2 < \mu_p^s.$$

Next we can choose $\tilde{\varepsilon}_1^2 = C_1 d^{-(p+1)s}$. Using Lemma 19, we can choose $C_1 < \mathfrak{C}_1^s$, where $\mathfrak{C}_1$ is the constant in Lemma 19, such that for any $d \geq \mathfrak{C}$, where $\mathfrak{C}$ is a constant only depending on s and p, we have

$$\tilde{\varepsilon}_1^2 < \mu_{p+1}^s.$$

Therefore, using Lemma 34 and Lemma 19, for any $d \geq \mathfrak{C}$, we have

$$\begin{aligned} V_2(\tilde{\varepsilon}_1, \mathcal{B}) \geq K(\tilde{\varepsilon}_1) &\geq \frac{1}{2} N(d, p+1) \ln\left(\frac{\mu_{p+1}^s}{\tilde{\varepsilon}_1^2}\right) \\ &\geq \frac{1}{2} N(d, p+1) \ln\left(\frac{\mathfrak{C}_1 d^{-(p+1)s}}{C_1 d^{-(p+1)s}}\right) \\ &= \frac{1}{2} N(d, p+1) \ln \frac{\mathfrak{C}_1}{C_1}. \end{aligned} \tag{101}$$

In addition, using Lemma 19 and Lemma 20, we have the following claim.

Claim 2 *Suppose that $\gamma \in (p+ps+s, (p+1)+(p+1)s]$ for some integer $p \geq 0$. Let $\tilde{\varepsilon}_2^2$ be defined as above. For any $\varepsilon_0 > 0$, there exists a sufficiently large constant $\mathfrak{C}$ only depending on s, p and ε_0, such that for any $d \geq \mathfrak{C}$, we have*

$$K\left(\sqrt{2}\sigma\tilde{\varepsilon}_2/6\right) \leq (1+\varepsilon_0)\frac{1}{2} N(d,p) \ln\left(\frac{18\mu_p^s}{\sigma^2\tilde{\varepsilon}_2^2}\right).$$

Therefore, for any $d \geq \mathfrak{C}$, where $\mathfrak{C}$ is a constant only depending on s, p and $\{a_j\}_{j\leq p+1}$,

$$\begin{aligned} V_K\left(\tilde{\varepsilon}_2, \mathcal{D}\right) = V_2\left(\sqrt{2}\sigma\tilde{\varepsilon}_2, \mathcal{B}\right) &\leq K\left(\frac{\sqrt{2}\sigma\tilde{\varepsilon}_2}{6}\right) \\ &\leq (1+\varepsilon_0)\,\frac{1}{2}N(d,p)\ln\left(\frac{18\mu_p^s}{\sigma^2\tilde{\varepsilon}_2^2}\right) \\ &\leq (1+\varepsilon_0)\,\frac{1}{2}N(d,p)\ln\left(\frac{18\mathfrak{C}_2 d^{-ps}}{\sigma^2 C_2 d^{-(p+1)s}}\right) \\ &= (1+\varepsilon_0)\,\frac{1}{2}N(d,p)\left(\ln\frac{18\mathfrak{C}_2}{\sigma^2 C_2} + s\ln d\right), \end{aligned} \tag{102}$$

where we use Lemma 33 and Lemma 34 for the first line and use Lemma 19 for the third line.

Using (98) and (99), also recalling that we assume $c_1 d^\gamma \leq n \leq c_2 d^\gamma$, we have

$$\frac{V_K(\tilde{\varepsilon}_2, \mathcal{D}) + n\tilde{\varepsilon}_2^2 + \ln 2}{V_2(\tilde{\varepsilon}_1, \mathcal{B})} \leq \frac{(1+\varepsilon_0)\,\frac{1}{2}N(d,p)\left(\ln\frac{18\mathfrak{C}_2}{\sigma^2 C_2} + s\ln d\right) + c_2 d^\gamma \cdot C_2 d^{-(p+1)s} + \ln 2}{\frac{1}{2}N(d,p+1)\ln\frac{\mathfrak{C}_1}{C_1}}. \tag{103}$$

Recalling that Lemma 21 shows $N(d,p) \leq \mathfrak{C}_4 d^p$ and $N(d,p+1) \geq \mathfrak{C}_3 d^{p+1}$ when $d \geq \mathfrak{C}$, the dominant terms in (103) are:

$$\frac{c_2 C_2 d^{\gamma-(p+1)s}}{\frac{1}{2}\ln\frac{\mathfrak{C}_1}{C_1}N(d,p+1)\ln d}.$$

Further noticing that $\gamma - (p+1)s \leq p+1$ for any $\gamma \in (p+ps+s, (p+1)+(p+1)s]$, so we can choose C_2 small enough and only depending on $s, \sigma, \gamma, \kappa, c_1, c_2$, such that

$$\frac{V_K(\tilde{\varepsilon}_2, \mathcal{D}) + n\tilde{\varepsilon}_2^2 + \ln 2}{V_2(\tilde{\varepsilon}_1, \mathcal{B})} \leq (103) := \mathfrak{c} < 1.$$

Then using Lemma 35 again, we have

$$\min_{\hat{f}} \max_{\rho_{f^*}\in\mathcal{D}} \mathbb{E}_{(\boldsymbol{X},\mathbf{y})\sim\rho_{f^*}^{\otimes n}} \left\|\hat{f} - f^*\right\|_{L^2}^2 \geq \frac{1-\mathfrak{c}}{4}\tilde{\varepsilon}_1^2 = \frac{1-\mathfrak{c}}{4}C_1 d^{-(p+1)s}.$$

Further recalling that $\mathcal{D} \subset \mathcal{P}$, we have

$$\min_{\hat{f}} \max_{\rho\in\mathcal{P}} \mathbb{E}_{(\boldsymbol{X},\mathbf{y})\sim\rho^{\otimes n}} \left\|\hat{f} - f_\rho^*\right\|_{L^2}^2 \geq \min_{\hat{f}} \max_{\rho_{f^*}\in\mathcal{D}} \mathbb{E}_{(\boldsymbol{X},\mathbf{y})\sim\rho_{f^*}^{\otimes n}} \left\|\hat{f} - f^*\right\|_{L^2}^2 \geq \frac{1-\mathfrak{c}}{4}C_1 d^{-(p+1)s}.$$

We finish the proof of Theorem 7 (ii).

■

Appendix D. Auxiliary results

The following proposition about estimating the L^2 norm with empirical norm is from Li et al. (2023a, Proposition C.9), which dates back to Caponnetto and Yao (2010).

Proposition 36 *Let μ be a probability measure on $\mathcal{X}$, $f \in L^2(\mathcal{X}, \mu)$ and $\|f\|_{L^\infty} \leq M$. Suppose we have $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ sampled i.i.d. from μ. Then for $\delta \in (0,1)$, the following holds with probability at least $1-\delta$:*

$$\frac{1}{2}\|f\|_{L^2}^2 - \frac{5M^2}{3n}\ln\frac{2}{\delta} \leq \|f\|_{L^2,n}^2 \leq \frac{3}{2}\|f\|_{L^2}^2 + \frac{5M^2}{3n}\ln\frac{2}{\delta}.$$

The following concentration inequality about self-adjoint Hilbert-Schmidt operator valued random variables is frequently used in related literature, e.g., Fischer and Steinwart (2020, Theorem 27) and Lin and Cevher (2020, Lemma 26).

Lemma 37 *Let $(\Omega, \mathcal{B}, P)$ be a probability space, $\mathcal{H}$ be a separable Hilbert space. Suppose that $A_1, \cdots, A_n$ are i.i.d. random variables with values in the set of self-adjoint Hilbert-Schmidt operators. If $\mathbb{E}A_i = 0$, and the operator norm $\|A_i\| \leq L, P$-a.e., and there exists a self-adjoint positive semi-definite trace class operator V with $\mathbb{E}A_i^2 \preceq V$. Then for $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$\left\|\frac{1}{n}\sum_{i=1}^n A_i\right\| \leq \frac{2L\beta}{3n} + \sqrt{\frac{2\|V\|\beta}{n}}, \quad \beta = \ln\frac{4\mathrm{trV}}{\delta\|V\|}.$$

The following Bernstein inequality about vector-valued random variables is frequently used, e.g., Caponnetto and de Vito (2007, Proposition 2) and Fischer and Steinwart (2020, Theorem 26).

Lemma 38 (Bernstein inequality) *Let $(\Omega, \mathcal{B}, P)$ be a probability space, H be a separable Hilbert space, and $\xi : \Omega \to H$ be a random variable with*

$$\mathbb{E}\|\xi\|_H^m \leq \frac{1}{2}m!\sigma^2 L^{m-2},$$

for all $m > 2$. Then for $\delta \in (0,1)$, ξ_i are i.i.d. random variables, with probability at least $1-\delta$, we have

$$\left\|\frac{1}{n}\sum_{i=1}^n \xi_i - \mathbb{E}\xi\right\|_H \leq 4\sqrt{2}\ln\frac{2}{\delta}\left(\frac{L}{n} + \frac{\sigma}{\sqrt{n}}\right).$$

Lemma 39 *Given the definition of $\mathcal{N}_1(\lambda)$ as in (4). If the condition (6) in Assumption 3 holds, we have*

$$\|T_\lambda^{-\frac{1}{2}}k(\boldsymbol{x}, \cdot)\|_{\mathcal{H}}^2 \leq \mathcal{N}_1(\lambda), \quad \mu\text{-a.e. } \boldsymbol{x} \in \mathcal{X}.$$

Proof

$$\begin{aligned}
\|T_\lambda^{-\frac{1}{2}}k(\boldsymbol{x}, \cdot)\|_{\mathcal{H}}^2 &= \Big\|\sum_{i=1}^\infty (\frac{1}{\lambda_i + \lambda})^{\frac{1}{2}}\lambda_i e_i(\boldsymbol{x})e_i(\cdot)\Big\|_{\mathcal{H}}^2 \\
&= \sum_{i=1}^\infty \frac{\lambda_i}{\lambda_i + \lambda}e_i^2(\boldsymbol{x}) \\
&\leq \mathcal{N}_1(\lambda), \quad \mu\text{-a.e. } \boldsymbol{x} \in \mathcal{X}.
\end{aligned}$$

■

Lemma 39 has a direct corollary.

Lemma 40 *Given the definition of $\mathcal{N}_1(\lambda)$ as in (4). If the condition (6) in Assumption 3 holds, we have*

$$\|T_\lambda^{-\frac{1}{2}} T_{\boldsymbol{x}} T_\lambda^{-\frac{1}{2}}\| \le \mathcal{N}_1(\lambda), \quad \mu\text{-a.e. } \boldsymbol{x} \in \mathcal{X}.$$

Proof Note that for any $f \in \mathcal{H}$,

$$\begin{aligned} T_\lambda^{-\frac{1}{2}} T_{\boldsymbol{x}} T_\lambda^{-\frac{1}{2}} f &= T_\lambda^{-\frac{1}{2}} K_{\boldsymbol{x}} K_{\boldsymbol{x}}^* T_\lambda^{-\frac{1}{2}} f \\ &= T_\lambda^{-\frac{1}{2}} K_{\boldsymbol{x}} \langle k(\boldsymbol{x}, \cdot), T_\lambda^{-\frac{1}{2}} f \rangle_{\mathcal{H}} \\ &= T_\lambda^{-\frac{1}{2}} K_{\boldsymbol{x}} \langle T_\lambda^{-\frac{1}{2}} k(\boldsymbol{x}, \cdot), f \rangle_{\mathcal{H}} \\ &= \langle T_\lambda^{-\frac{1}{2}} k(\boldsymbol{x}, \cdot), f \rangle_{\mathcal{H}} \cdot T_\lambda^{-\frac{1}{2}} k(\boldsymbol{x}, \cdot). \end{aligned}$$

So $\|T_\lambda^{-\frac{1}{2}} T_{\boldsymbol{x}} T_\lambda^{-\frac{1}{2}}\| = \sup_{\|f\|_{\mathcal{H}}=1} \|T_\lambda^{-\frac{1}{2}} T_{\boldsymbol{x}} T_\lambda^{-\frac{1}{2}} f\|_{\mathcal{H}} = \sup_{\|f\|_{\mathcal{H}}=1} \langle T_\lambda^{-\frac{1}{2}} k(\boldsymbol{x}, \cdot), f \rangle_{\mathcal{H}} \cdot \|T_\lambda^{-\frac{1}{2}} k(\boldsymbol{x}, \cdot)\|_{\mathcal{H}} = \|T_\lambda^{-\frac{1}{2}} k(\boldsymbol{x}, \cdot)\|_{\mathcal{H}}^2$. Using Lemma 39, we finish the proof. ■

We state the following three lemmas without proof, since the proofs are classical and the verification about the constants is tedious. We refer to Appendix A in Zhang et al. (2024) and the references therein for the proof, the definition of *Lorentz space* $L^{p,q}(\mathcal{X}, \mu)$ and *real interpolation* $(\cdot, \cdot)_{\theta,q}$.

Lemma 41 *Let μ be the probability distribution on $\mathcal{X}$. For $1 < p_1 \neq p_2 < \infty$, $1 \le q \le \infty$ and $0 < \theta < 1$, we have*

$$(L^{p_1}(\mathcal{X}, \mu), L^{p_2}(\mathcal{X}, \mu))_{\theta,q} \cong L^{p_\theta, q}(\mathcal{X}, \mu), \quad \frac{1}{p_\theta} = \frac{1-\theta}{p_1} + \frac{\theta}{p_2},$$

where $L^{p_\theta,q}(\mathcal{X}, \mu)$ is the Lorentz space and the equivalent norm only involves absolute constants.

Lemma 42 *Let μ be the probability distribution on $\mathcal{X}$. If $1 < p < \infty$ and $1 \le q_1 \le q_2 \le \infty$, we have*

$$L^{p,q_1}(\mathcal{X}, \mu) \hookrightarrow L^{p,q_2}(\mathcal{X}, \mu),$$

and the operator norm are upper bounded by an absolute constant.

Lemma 43 *Let μ be the probability distribution on $\mathcal{X}$. For $1 < p < \infty$, we have*

$$L^{p,p}(\mathcal{X}, \mu) \cong L^p(\mathcal{X}, \mu); \quad L^{p,\infty}(\mathcal{X}, \mu) \cong L^{p,w}(\mathcal{X}, \mu),$$

where $L^{p,w}(\mathcal{X}, \mu)$ denotes the weak L^p space and the equivalent norm only involves absolute constants.

Theorem 44 (L^q-embedding property) *Suppose that $\mathcal{H}$ is the RKHS associated with a continuous, positive-definite and symmetric kernel k on a compact set $\mathcal{X} \subset \mathbb{R}^d$ and the probability distribution on $\mathcal{X}$ is μ. Further suppose that $\sup_{\boldsymbol{x}\in\mathcal{X}} |k(x,x)| \leq \kappa^2$, where κ is an absolute constant. Then for any $0 < s < 1$, we have*

$$[\mathcal{H}]^s \hookrightarrow L^{q_s}(\mathcal{X},\mu), \quad \forall q_s < \frac{2}{1-s},$$

and there exists a constant $C_{s,\kappa}$ only depending on s and κ, such that the operator norm of the embedding operator satisfies

$$\|[\mathcal{H}]^s \hookrightarrow L^{q_s}(\mathcal{X},\mu)\| \leq C_{s,\kappa}.$$

Proof Denote $(E_0, E_1)_{\theta,q}$ as the real interpolation of two normed spaces. Steinwart and Scovel (2012, Theorem 4.6) shows that for $0 < s < 1$,

$$[\mathcal{H}]^s \cong \left(L^2(\mathcal{X},\mu), [\mathcal{H}]^1\right)_{s,2},$$

where the equivalent norm involves constants only depending on s.

Since $\sup_{\boldsymbol{x}\in\mathcal{X}} |k(\boldsymbol{x},\boldsymbol{x})| \leq \kappa^2$ implies that the operator norm of embedding $I_1 : \mathcal{H} \hookrightarrow L^\infty$ satisfies $\|I_1\| \leq \kappa^2$. Define $I_2 : \left(L^2, \mathcal{H}\right)_{\theta,2} \hookrightarrow \left(L^2, L^\infty\right)_{\theta,2}$ for some $\theta \in (0,1)$, the definition of real interpolation through *K-Method* (see Chapter 22 in Tartar 2007) actually implies $\|I_2\| \leq \max\{1, \kappa^2\}$. Then any $0 < M < \infty$, using Lemma 41, we have

$$[\mathcal{H}]^s \hookrightarrow \left(L^2(\mathcal{X},\mu), L^M(\mathcal{X},\mu)\right)_{s,2} \cong L^{q_s',2}(\mathcal{X},\mu),$$

where $\frac{1}{q_s'} = \frac{1-s}{2} + \frac{s}{M}$.

For any $q_s < \frac{2}{1-s}$, we can choose M large enough such that $q_s' > q_s$. Further, since $0 < s < 1$ and thus $q_s' > q_s > 2$, using Lemma 42 and Lemma 43, we have

$$L^{q_s',2}(\mathcal{X},\mu) \hookrightarrow L^{q_s',q_s'}(\mathcal{X},\mu) \cong L^{q_s'}(\mathcal{X},\mu) \hookrightarrow L^{q_s}(\mathcal{X},\mu).$$

We finish the proof. ■

In the following, we provide a remark on Theorem 44.

Remark 45 *Intuitively, there should be a constant depending on the dimension d in the operator norm of the embedding. For instance, their are extensive literature studying the dependence of the embedding constants on d in the Sobolev type inequalities (Cotsiolis and Tavoularis, 2004; Mizuguchi et al., 2016; Novak et al., 2018).*

Denote $(I-\Delta)^{-\frac{r}{2}}, r > 0$, as the Bessel potential operators (see, e.g., Section 2 of Cotsiolis and Tavoularis 2004). Then the fractional Sobolev space can be defined as $H^r(\mathbb{R}^d) = \left\{ f \in L^2(\mathbb{R}^d) \;\middle|\; \|(I-\Delta)^{\frac{r}{2}} f\|_{L^2} < \infty \right\}$, $r > 0$. It is well known that when $r > \frac{d}{2}$, $H^r(\mathbb{R}^d)$ is an RKHS with bounded kernel function. When $r < \frac{d}{2}$, denote $I_{d,r}$ as the embedding from $H^r(\mathbb{R}^d)$ to $L^{\frac{2d}{d-2r}}(\mathbb{R}^d)$. Theorem 1.1 in Cotsiolis and Tavoularis (2004) gives an upper bound of the

operator norm $\|I_{d,r}\|$ *for any* $d > 0, 0 < r < \frac{d}{2}$ *(note that* $\|(-\Delta)^{\frac{r}{2}} f\|_{L^2} \leq \|(I-\Delta)^{\frac{r}{2}} f\|_{L^2}$*). Since for* $0 < s < 1$, $[H^{\frac{d}{2}}(\mathbb{R}^d)]^s \cong (L^2(\mathbb{R}^d), H^{\frac{d}{2}}(\mathbb{R}^d))_{s,2} \cong H^{\frac{ds}{2}}(\mathbb{R}^d))$, *letting* $r = \frac{d}{2}$, *we have*

$$H^r(\mathbb{R}^d) = [\mathcal{H}_d]^{\frac{2}{3}}, \quad \textit{where} \quad \mathcal{H}_d = H^{\frac{d}{2}}(\mathbb{R}^d) \quad \textit{is an RKHS.}$$

(With a little abusement of notation, we consider $H^{\frac{d}{2}}(\mathbb{R}^d)$ *as an RKHS). Then* $I_{d,\frac{d}{2}}$ *can also be interpreted as*

$$I_{d,\frac{d}{2}} : [\mathcal{H}_d]^s \hookrightarrow L^{\frac{2}{1-s}}(\mathbb{R}^d), \ \textit{with } s = \frac{2}{3}.$$

Detailed calculation about the constant in Theorem 1.1 in Cotsiolis and Tavoularis (2004) shows that the operator norm of $I_{d,\frac{d}{2}}$ *decreases to 0, i.e.,*

$$\|I_{d,\frac{d}{2}}\| \to 0, \quad \textit{as} \quad d \to \infty.$$

This indicates that although the embedding norm may depend on d, it can always be upper bounded by a constant. This shows the consistency with Theorem 44 in our paper.

So when will the embedding norm tend to 0 and when will it remain as a constant? We can get some inspiration from the operator norm of $I_1 : \mathcal{H}_d \hookrightarrow L^\infty$*. Recall that we assume* $\sup_{\boldsymbol{x} \in \mathcal{X}} |k(\boldsymbol{x}, \boldsymbol{x})| \leq \kappa^2$ *in Theorem 44, where* κ *is an absolute constant. This directly implies* $\|I_1\| \leq \kappa^2$*. This assumption is appropriate for some RKHSs, for instance, the inner product kernel in Assumption 4 and NTK on the sphere. For these RKHSs,* $\sup_{\boldsymbol{x} \in \mathcal{X}} |k(x,x)| \leq \kappa^2$ *will not change as* $d \to \infty$*. However, we conjecture that the bound* $\|I_1\| \leq \kappa^2$ *may be too loose for other RKHSs when* $d \to \infty$*.*

Let us see the example of fractional Sobolev space again. For $d, r \in \mathbb{N}$ *with* $r > \frac{d}{2}$*, denote* $I_{d,r,\infty}$ *as the embedding from* $H^r(\mathbb{R}^d)$ *to* $L^\infty(\mathbb{R}^d)$*. Theorem 11 in Novak et al. (2018) shows that* $\|I_{d,r,\infty}\| \to 0$ *as* $d \to \infty$*. Note that the definition of* $H^r(\mathbb{R}^d)$ *in this section is actually the same as the definition in Section 4.1 of Novak et al. (2018).*

References

Michael Aerni, Marco Milanta, Konstantin Donhauser, and Fanny Yang. Strong inductive biases provably prevent harmless interpolation. In *The Eleventh International Conference on Learning Representations*, 2022.

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. *Advances in Neural Information Processing Systems*, 32, 2019.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

F. Bauer, S. Pereverzyev, and L. Rosasco. On regularization algorithms in learning theory. *Journal of complexity*, 23(1):52–72, 2007.

Daniel Beaglehole, Mikhail Belkin, and Parthe Pandit. On the inconsistency of kernel ridgeless regression in fixed dimensions. *SIAM Journal on Mathematics of Data Science*, 5(4):854–872, 2023.

Alberto Bietti and Julien Mairal. On the inductive bias of neural tangent kernels. *Advances in Neural Information Processing Systems*, 32, 2019.

Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *International Conference on Machine Learning*, pages 1024–1034. PMLR, 2020.

Simon Buchholz. Kernel interpolation in Sobolev spaces is not consistent in low dimensions. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 3410–3440. PMLR, July 2022.

Andrea Caponnetto. Optimal rates for regularization operators in learning theory. Technical report, MASSACHUSETTS INST OF TECH CAMBRIDGE COMPUTER SCIENCE AND ARTIFICIAL ..., 2006.

Andrea Caponnetto and Ernesto de Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7:331–368, 2007.

Andrea Caponnetto and Yuan Yao. Cross-validation based adaptation for regularization operators in learning theory. *Analysis and Applications*, 8(02):161–183, 2010.

Athanase Cotsiolis and Nikolaos K Tavoularis. Best constants for sobolev inequalities for higher order fractional derivatives. *Journal of mathematical analysis and applications*, 295 (1):225–236, 2004.

Hugo Cui, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime. *Advances in Neural Information Processing Systems*, 34:10131–10143, 2021.

Feng Dai and Yuan Xu. *Approximation Theory and Harmonic Analysis on Spheres and Balls*. Springer Monographs in Mathematics. Springer New York, New York, NY, 2013. ISBN 978-1-4614-6659-8 978-1-4614-6660-4. doi: 10.1007/978-1-4614-6660-4.

Konstantin Donhauser, Mingqi Wu, and Fanny Yang. How rotational invariance of common kernels prevents generalization in high dimensions. In *International Conference on Machine Learning*, pages 2804–2814. PMLR, 2021.

Simon-Raphael Fischer and Ingo Steinwart. Sobolev norm learning rates for regularized least-squares algorithms. *Journal of Machine Learning Research*, 21:205:1–205:38, 2020.

Jean Gallier, Jocelyn Quaintance, Jean Gallier, and Jocelyn Quaintance. Spherical harmonics and linear representations of lie groups. *Differential Geometry and Lie Groups: A Second Course*, pages 265–360, 2020.

L. Lo Gerfo, Lorenzo Rosasco, Francesca Odone, E. De Vito, and Alessandro Verri. Spectral algorithms for supervised learning. *Neural Computation*, 20(7):1873–1897, 2008.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. When do neural networks outperform kernel methods? *Advances in Neural Information Processing Systems*, 33:14820–14830, 2020.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *The Annals of Statistics*, 49(2):1029 – 1054, 2021. doi: 10.1214/20-AOS1990. URL https://doi.org/10.1214/20-AOS1990.

Nikhil Ghosh, Song Mei, and Bin Yu. The three stages of learning dynamics in high-dimensional kernel methods. In *International Conference on Learning Representations*, 2021.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949 – 986, 2022. doi: 10.1214/21-AOS2133. URL https://doi.org/10.1214/21-AOS2133.

Hong Hu and Yue M Lu. Sharp asymptotics of kernel ridge regression beyond the linear regime. *arXiv preprint arXiv:2205.06798*, 2022.

Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Noureddine El Karoui. The spectrum of kernel random matrices. *The Annals of Statistics*, 38 (1):1 – 50, 2010. doi: 10.1214/08-AOS648. URL https://doi.org/10.1214/08-AOS648.

Jianfa Lai, Manyun Xu, Rui Chen, and Qian Lin. Generalization ability of wide neural networks on $\mathbb{R}$. *arXiv preprint arXiv:2302.05933*, 2023.

Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in neural information processing systems*, 32, 2019.

Yicheng Li, Haobo Zhang, and Qian Lin. On the saturation effect of kernel ridge regression. In *International Conference on Learning Representations*, February 2023a.

Yicheng Li, Haobo Zhang, and Qian Lin. On the asymptotic learning curves of kernel ridge regression under power-law decay. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b.

Yicheng Li, Haobo Zhang, and Qian Lin. Kernel interpolation generalizes poorly. *Biometrika*, 111(2):715–722, 2024.

Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel "Ridgeless" regression can generalize. *The Annals of Statistics*, 48(3):1329 – 1347, 2020. doi: 10.1214/19-AOS1849. URL https://doi.org/10.1214/19-AOS1849.

Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai. On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In *Conference on Learning Theory*, pages 2683–2711. PMLR, 2020.

Junhong Lin and Volkan Cevher. Optimal convergence for distributed learning with stochastic gradient methods and spectral algorithms. *Journal of Machine Learning Research*, 21: 147–1, 2020.

Junhong Lin, Alessandro Rudi, L. Rosasco, and V. Cevher. Optimal rates for spectral algorithms with least-squares regression over Hilbert spaces. *Applied and Computational Harmonic Analysis*, 48:868–890, 2018.

Fanghui Liu, Zhenyu Liao, and Johan Suykens. Kernel regression in high dimensions: Refined analysis beyond double descent. In *International Conference on Artificial Intelligence and Statistics*, pages 649–657. PMLR, 2021.

Weihao Lu, Haobo Zhang, Yicheng Li, Manyun Xu, and Qian Lin. Optimal rate of kernel regression in large dimensions. *arXiv preprint arXiv:2309.04268*, 2023.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Generalization error of random feature and kernel methods: Hypercontractivity and kernel matrix concentration. *Applied and Computational Harmonic Analysis*, 59:3–84, 2022.

Theodor Misiakiewicz. Spectrum of inner-product kernel matrices in the polynomial regime and multiple descent phenomenon in kernel ridge regression. *arXiv preprint arXiv:2204.10425*, 2022.

Makoto Mizuguchi, Akitoshi Takayasu, Takayuki Kubo, and Shin'ichi Oishi. On the embedding constant of the sobolev type inequality for fractional derivatives. *Nonlinear Theory and Its Applications, IEICE*, 7(3):386–394, 2016.

Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 1(1):67–83, 2020. doi: 10.1109/JSAIT.2020.2984716.

Erich Novak, Mario Ullrich, Henryk Woźniakowski, and Shun Zhang. Reproducing kernels of sobolev spaces on r d and applications to embedding constants and tractability. *Analysis and Applications*, 16(05):693–715, 2018.

Alexander Rakhlin and Xiyu Zhai. Consistency of interpolation with laplace kernels is a high-dimensional phenomenon. In *Conference on Learning Theory*, pages 2595–2623. PMLR, 2019.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Early stopping and non-parametric regression: an optimal data-dependent stopping rule. *The Journal of Machine Learning Research*, 15(1):335–366, 2014.

Ingo Steinwart and Andreas Christmann. Support vector machines. In *Information Science and Statistics*, 2008.

Ingo Steinwart and C. Scovel. Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs. *Constructive Approximation*, 35(3):363–417, 2012.

Ingo Steinwart, D. Hush, and C. Scovel. Optimal rates for regularized least squares regression. In *COLT*, pages 79–93, 2009.

Luc Tartar. *An introduction to Sobolev spaces and interpolation spaces*, volume 3. Springer Science & Business Media, 2007.

Alexander Tsigler and Peter L Bartlett. Benign overfitting in ridge regression. *Journal of Machine Learning Research*, 24(123):1–76, 2023.

L Xiao, H Hu, T Misiakiewicz, Y Lu, and J Pennington. Precise learning curves and higher-order scaling limits for dot product kernel regression. In *Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564 – 1599, 1999. doi: 10.1214/aos/1017939142. URL `https://doi.org/10.1214/aos/1017939142`.

Haobo Zhang, Yicheng Li, Weihao Lu, and Qian Lin. On the optimality of misspecified kernel ridge regression. In *International Conference on Machine Learning*, pages 41331–41353. PMLR, 2023.

Haobo Zhang, Yicheng Li, and Qian Lin. On the optimality of misspecified spectral algorithms. *Journal of Machine Learning Research*, 25(188):1–50, 2024.